Text analysis 1

Lukáš Lehotský

Why text analysis?

Text as discourse

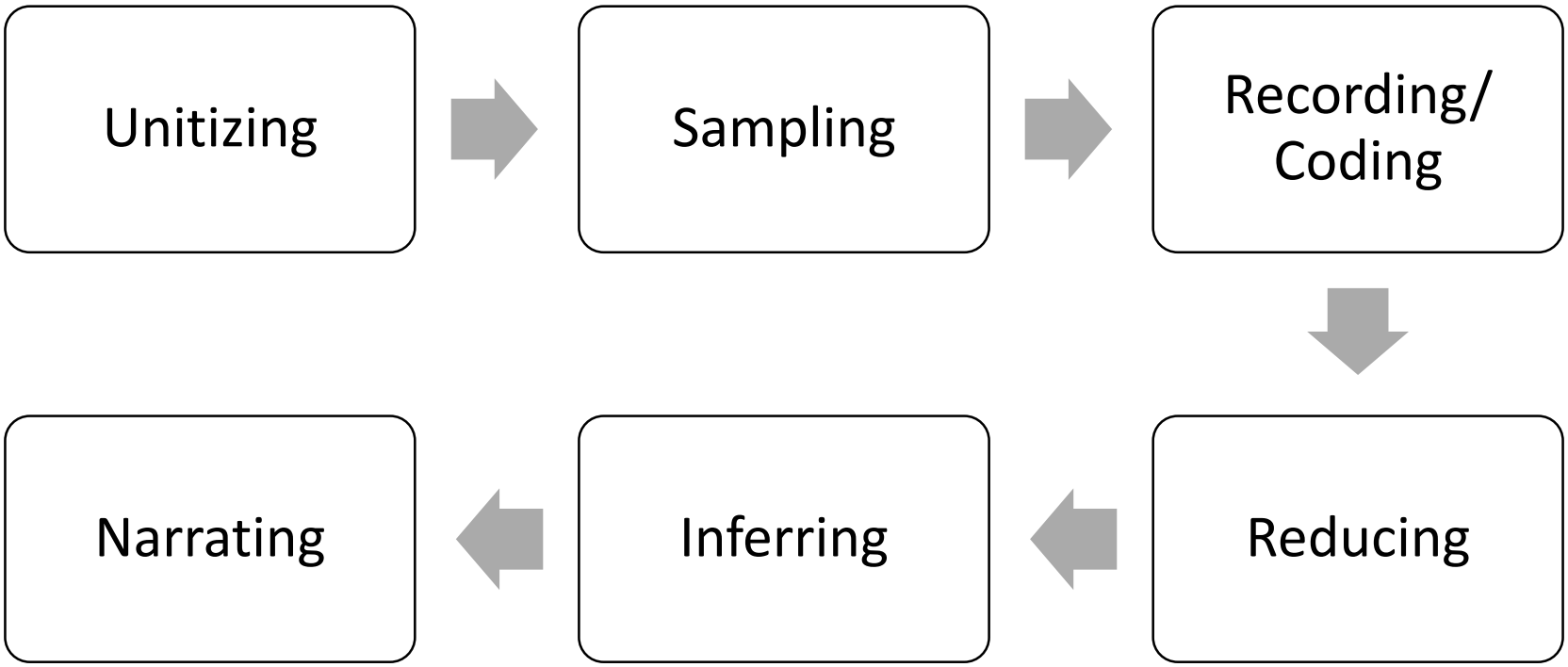
Text as patterns

Manifest vs. latent content

“text analysis is just a fancy and convoluted way how to obtain independent or dependent variable”

Inaki Sagarzazu 2016

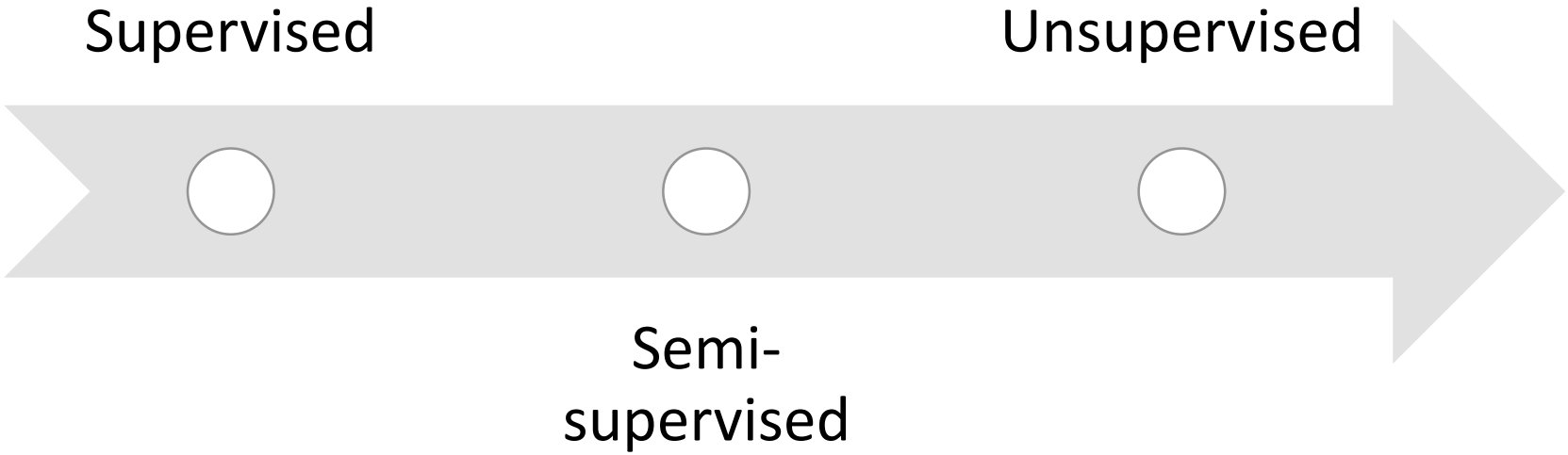
Design of CA research

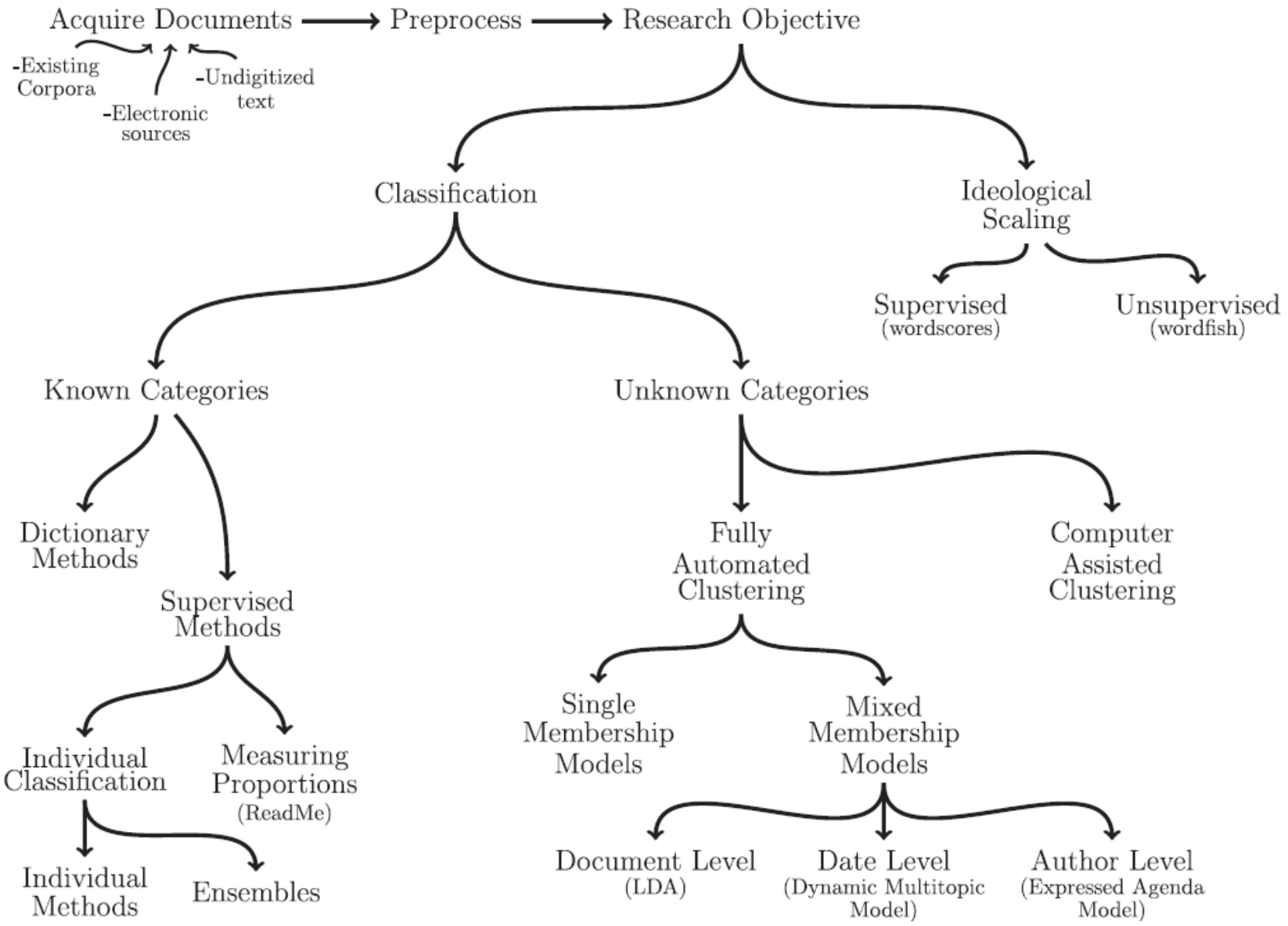


Krippendorff 2013, p. 86

Methods

Methods of TA





(Grimmer & Steweart 2013)

Methods of text analysis

- Supervised methods
  - Kewords in context (KWIC)
  - Manual coding
- Semi-supervised
  - Dictionary-based methods
    - Deductively given dictionary
    - Dictionary obtained from data
      - Automatically
      - Manually
- Unsupervised
  - Frequencies
  - Topic modeling
  - …

Fully supervised

- Requires manual text processing
- Most approaches based on manual coding of text units
- Inductive vs. deductive coding
  - Inductive – data-driven
    - Categories not known
    - Open coding – categories emerge in iterative text reading
    - Axial coding – abstraction from open coding into categories
    - Constant comparative approach – re-reading already coded units
  - Deductive – theory-driven
    - Categories known a-priori
    - Existing code-book applied over data
    - Comparative Manifesto Project

Keywords in Context (KWIC)

- Relational analysis of a concept
  - Analyzes context of the concept through the way the concept is used within the text – **original linguistic environment** of the text
- Exploration of the corpus
- Requires prior knowledge of keywords
- Input for further analysis
  - Dictionary construction
  - Frequency analysis
  - Coding

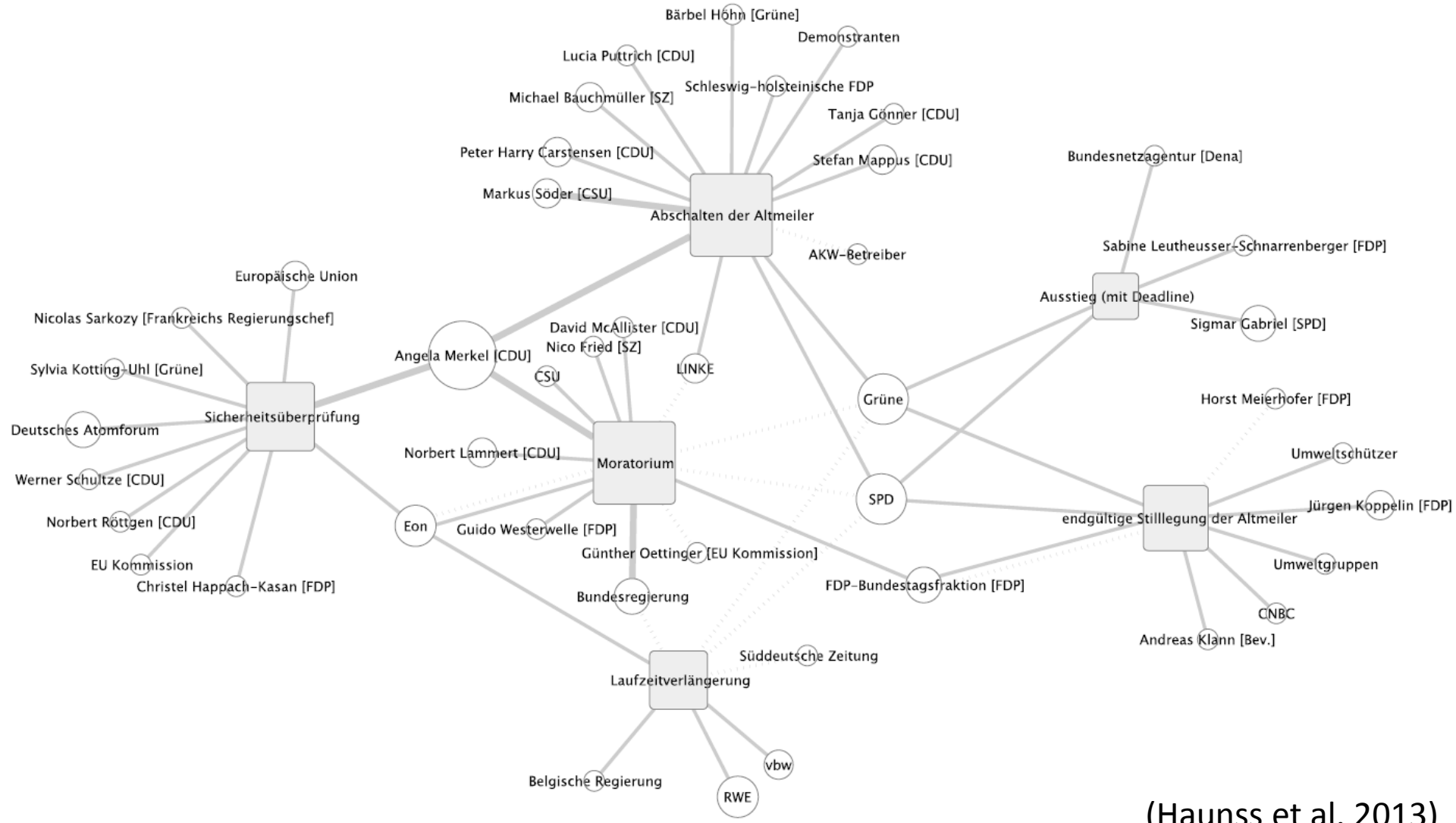
Keywords in Context (KWIC)

exchange of information about	**energy**	policy and coordination of
and coordination of the	**energy**	policy of V 4
sphere of new EU	**energy**	legislation, especially rule
of trans- European	**Energy**	Networks, concentrate on
with the operation of	**energy**	facility, impact of
the field of the	**energy**	sector, industry and
	Energy	continuation of meeting of
establishment of a common	**energy**	and gas market.
- operation in the	**energy**	sector in the usual

Fully supervised

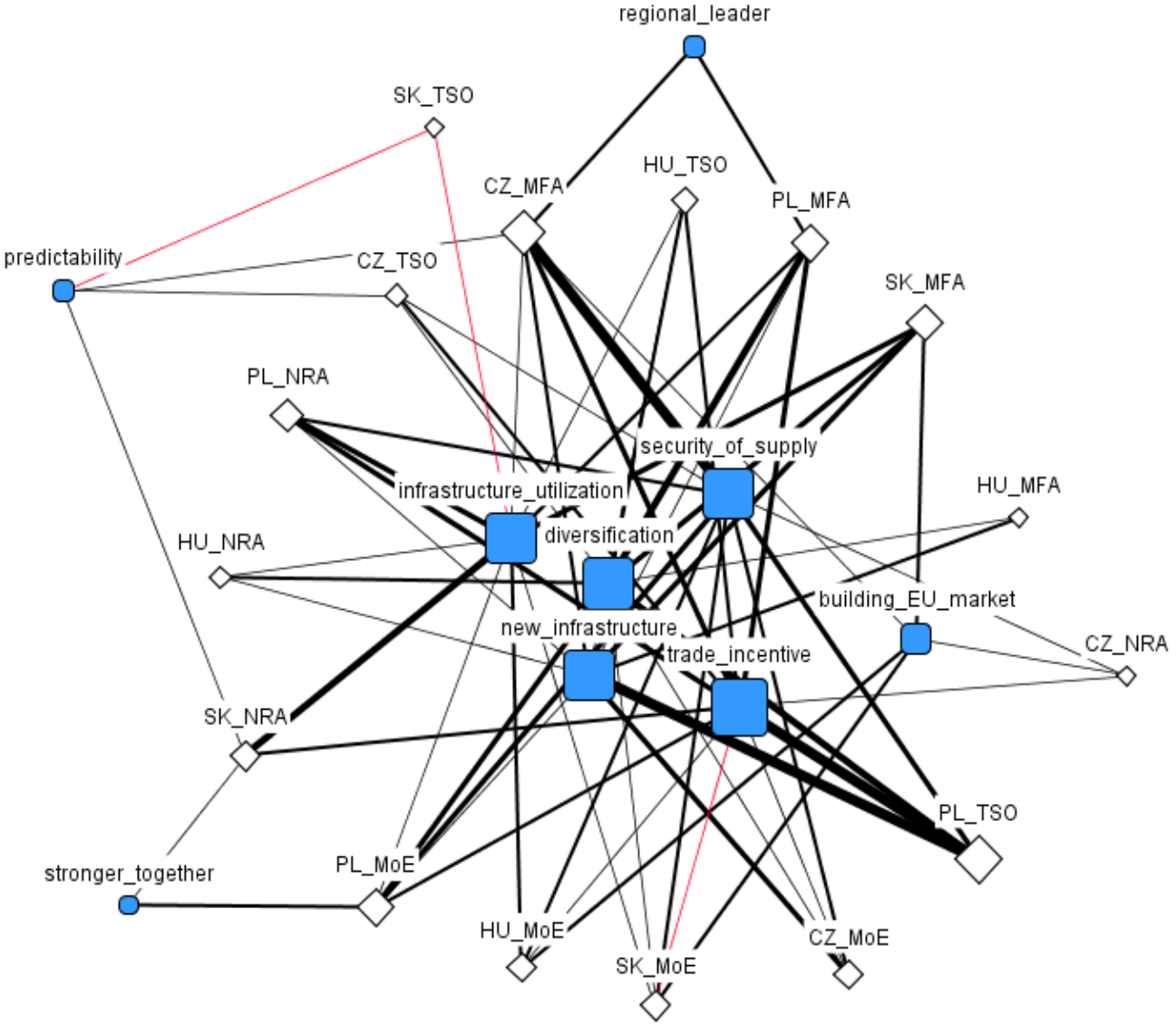
- Discourse analysis
- Socio-semantic networks
- Discourse network analysis
  - Socio-semantic networks of actors and meanings (codes) they use

Fully supervised - DNA



(Haunss et al. 2013)

Fully supervised - DNA



Issues with manual coding

- High validity, questionable reliability
- Reliability of human coders needs to be measured and accounted for
  - Intra-coder reliability (variation by same coder)
  - Inter-coder reliability (variation by different coder)
- Krippendorff α, Kohen κ
- Disagreements need to be solved
  - Only overlap (strict)
  - Resolution of differences

http://bit.ly/meb421coding

Semi-supervised

- Dictionary-based automated coding
  - Words in dictionary are discovered across the corpus
  - Coding process is done automatically
- Construction of dictionaries
  - Given pre-defined dictionary
    - WordStat, LIWC, …
  - Constructed from data
    - Theoretically-informed
    - Automatically generated
      - WordScores
      - Wordfish

Semi-supervised

- Existing dictionaries
  - WordStat (Laver & Garry 2000)
    - Estimation of policy positions from political texts
    - 415 words, 19 categories
  - LIWC (Linguistic Inquiry Word Count)
  - Sentiment dictionaries
  - General Inquirer
- Logic of this approach is to crawl over texts, discover tokens in dictionary and score texts
  - Scoring whole corpus
  - Scoring individual texts

Semi-supervised

- Dictionary from data
  - Sample of texts with known properties
  - Other texts related – e.g. legal/conceptual documents
- Dictionaries built by researchers
  - Long process
  - High validity – researchers know texts
  - Lower reliability – same reasons as manual coding

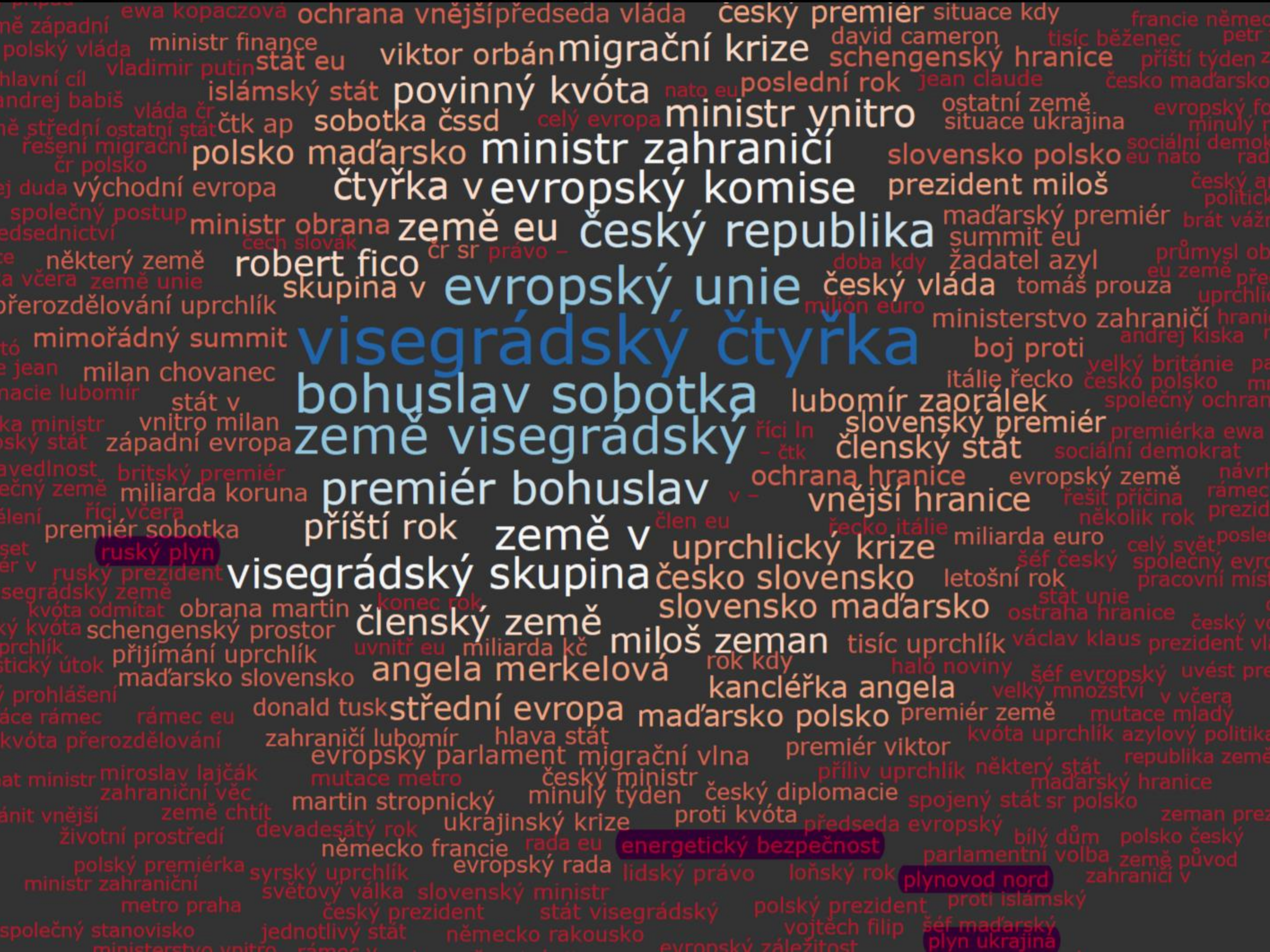
Semi-supervised

- WordScores - automated dictionary constructing
  - Laver, Benoit and Garry 2003
  - Two populations of texts
    - Texts with known properties – training set
    - Texts with unknown properties – target set
  - Logic of the process
    - Assign values of the category to known texts (training sample)
    - Let computer find words in the training sample and assign individual scores to words from texts
    - Code unknown texts with the generated dictionary
  - High reliability, but questionable validity
  - Single dimension
  - Word meanings have to be stable over time

Unsupervised

- Most naïve – word frequencies
  - Just a crude exploratory hint of what is in texts
- Clustering and multidimensional scaling of words
  - Based on co-occurrence of words
- Correspondence analysis
- Unsupervised scaling
- Unsupervised categorization on term-document matrix
  - Topic modeling
- Co-occurrence term networks

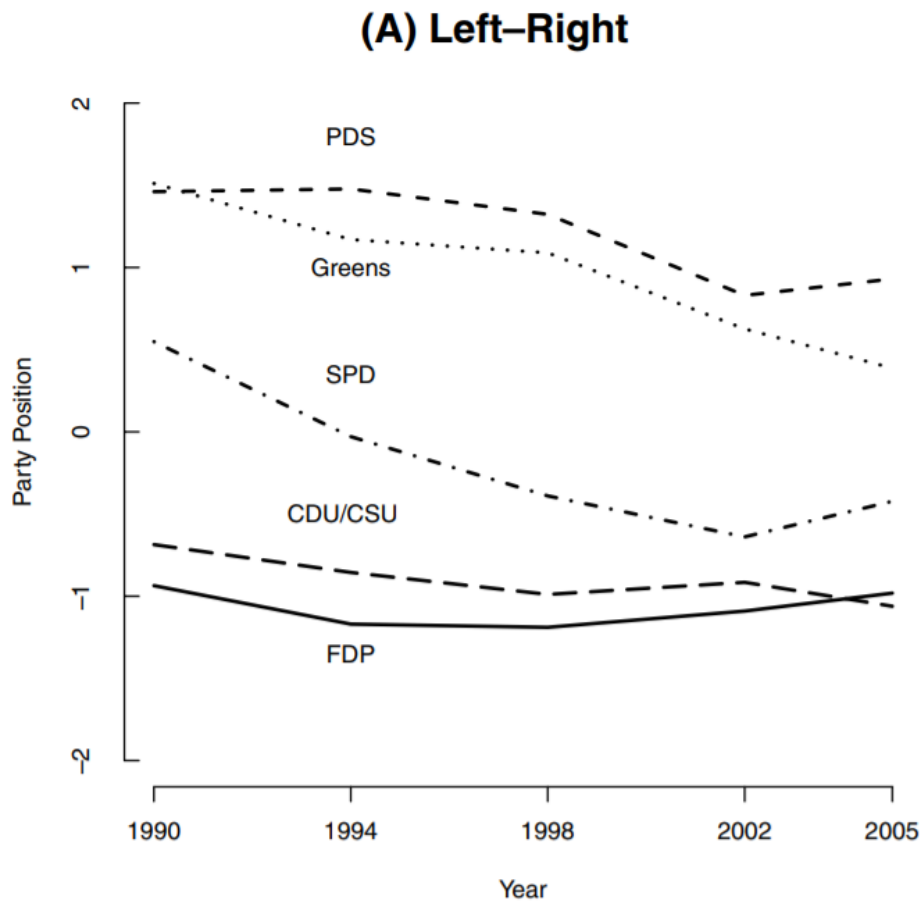

prostředek
pomáhat
martin
blízký
konec
schengen
festival
existovat
jediný
znamenat
vlna
úřad
národní
důležitý
solidarita
viktor
bezpečnost
práce
příští
stále
orbán
takzvaný
mimo
dítě
dohodnout
diskuse
kdyby
rozhodnout
mezinárodní
souvislost
hovořit
fórum
debata
unijní
uprchlický
vztah
myslit
volba
řada
především
často
významný
nejen
zákon
konference
přijmout
obrana
případ
poslední
zástupce
poslanec
platit
brzy
občan
trh
některý
zahraniční
prostor
návštěva
czech
cesta
rusko
zeman
právo
vůči
kvůli
vysoký
rozdíl
brát
čas
autor
dohoda
sám
jednat
kontrola
pomoc
skupina
visegrádský
postoj
téma
shodnout
zaorálek
summit
maďarsko
maďarský
současný
patřit
vojenský
visegrád
republika
postup
hlava
čekat
procento
společný
kvóta
hranice
návrh
ministerstvo
divadlo
spíše
totiž
nějaký
řešení
vědět
sobotka
politika
ekonomický
přístup
klíčový
milión
řešit
ruský
tlak
jít
místo
uprchlík
zahraničí
výsledek
ovšem
německo
vláda
evropa
krize
energetický
život
chovanec
miliarda
čtyřka
fond
přitom
peníze
lze
slovenský
oba
člověk
usa
zájem
doba
důvod
migrační
pan
stát
premiér
tam
názor
schengenský
pozice
merkelová
polsko
takový
říkat
západní
tady
spolupráce
věc
česko
kolega
chtít
možnost
schůzka
informace
kancléřka
ukrajinský
včera
situace
region
začít
síla
moc
země
například
zahrada
připravit
dostat
proti
český
rok
ukrajina
udělat
služba
čtk
program
praha
plyn
evropský
šéf
otázka
podpora
město
americký
slovensko
stejně
euro
vidět
jednání
velmi
student
ministr
říci
velký
uvést
dodat
svět
zároveň
datetimestamp
unie
muset
nato
povinný
jasný
cena
slovák
zůstat
oblast
běženec
zdůraznit
německý
angela
východní
problém
tisíc
dát
prezident
akce
řecko
opatření
daleko
snažit
krok
projekt
politický
vnitro
možný
policista
odmítat
měsíc
tiskový
nyní
rámec
polský
migrant
společnost
většina
malý
západ
afrika
podpořit
část
předseda
bohuslav
komise
fico
prohlásit
turecko
británie
miloš
druhý
kdy
člen
dobrý
brusel
celý
mluvit
změna
základ
rakouský
jan
několik
vnější
válka
francie
naopak
veřejný
minulý
zejména
žádný
ochrana
jaký
rád
rakousko
podporovat
demokracie
kyjev
pozice
odmítnout
přijít
sociální
zdroj
málo
armáda
nikdo
centrum
mluvčí
letošní
boj
rada
týden
dále
robert

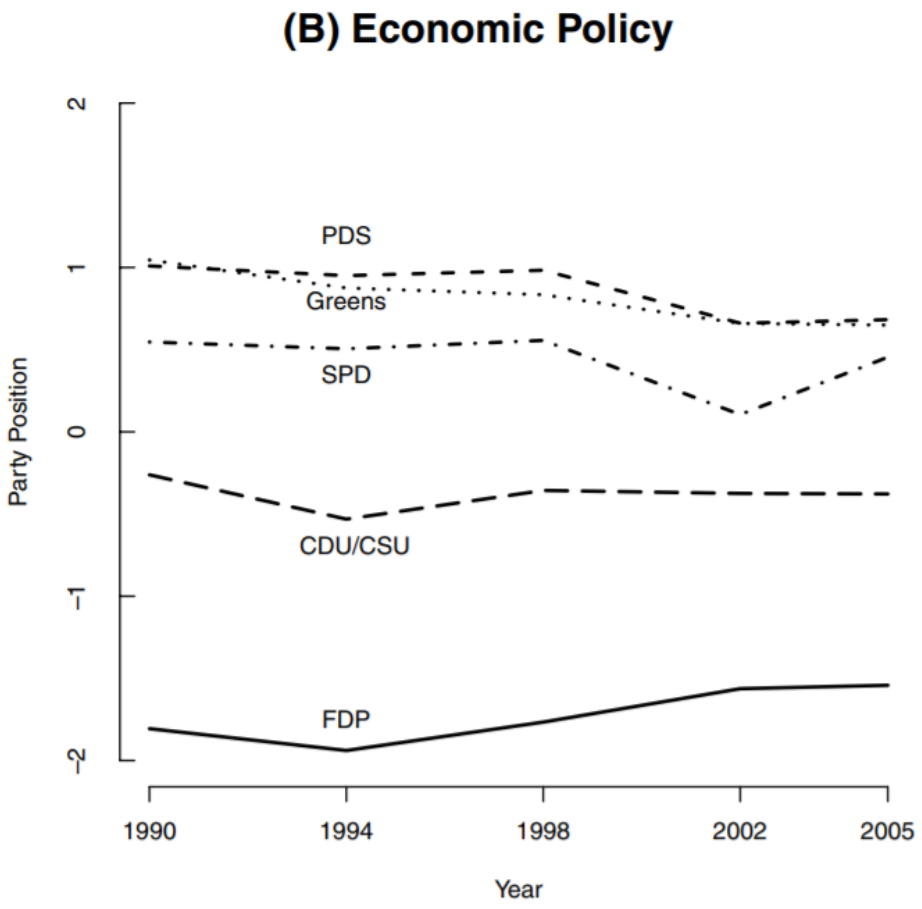

visegrádský čtyřka
bohuslav sobotka
země visegrádský
evropský unie
český republika
premiér bohuslav
evropský komise
ministr zahraničí
ministr vnitro
povinný kvóta
migrační krize
visegrádský skupina
země v
členský země
miloš zeman
robert fico
polsko maďarsko
čtyřka v
země eu
příští rok
angela merkelová
střední evropa
sobotka čssd
viktor orbán
islámský stát
ochrana vnější
předseda vláda
český premiér
schengenský hranice
poslední rok
slovensko polsko
prezident miloš
maďarský premiér
summit eu
žadatel azyl
český vláda
ministerstvo zahraničí
boj proti
lubomír zaorálek
slovenský premiér
členský stát
ochrana hranice
vnější hranice
uprchlický krize
česko slovensko
slovensko maďarsko
tisíc uprchlík
kancléřka angela
maďarsko polsko
východní evropa
ministr obrana
skupina v
mimořádný summit
milan chovanec
stát v
vnitro milan
západní evropa
miliarda koruna
premiér sobotka
ruský plyn
schengenský prostor
obrana martin
přijímání uprchlík
maďarsko slovensko
donald tusk
zahraničí lubomír
hlava stát
evropský parlament
migrační vlna
český ministr
minulý týden
český diplomacie
proti kvóta
martin stropnický
ukrajinský krize
německo francie
evropský rada
energetický bezpečnost
plynovod nord
premiér viktor
ewa kopaczová
ministr finance
stát eu
vladimir putin
david cameron
tisíc běženec
jean claude
ostatní země
situace ukrajina
situace kdy
ministr zahraniční
miroslav lajčák
zahraniční věc
země chtít
devadesátý rok
polský premiérka
syrský uprchlík
světový válka
slovenský ministr
metro praha
český prezident
stát visegrádský
společný stanovisko
jednotlivý stát
německo rakousko
polský prezident
vojtěch filip
šéf maďarský
plyn ukrajina
proti islámský
lidský právo
loňský rok
parlamentní volba
bílý dům
polsko český
země původ
předseda evropský
spojený stát
sr polsko
maďarský hranice
republika země
některý stát
příliv uprchlík
kvóta uprchlík
azylový politika
premiér země
mutace mladý
velký množství
šéf evropský
halo noviny
václav klaus
ostraha hranice
stát unie
letošní rok
šéf český
miliarda euro
celý svět
pracovní místo
několik rok
řešit příčina
evropský země
sociální demokrat
premiérka ewa
společný ochrana
česko polsko
velký británie
itálie řecko
andrej kiska
tomáš prouza
milión euro
doba kdy
čech slovák
čr sr
právo
některý země
země unie
přerozdělování uprchlík
společný postup
čr polsko
řešení migrační
vláda čr
čtk ap
celý evropa
nato eu
hlavní cíl
andrej babiš
polský vláda
britský premiér
říci včera
ruský prezident
kvóta odmítat
konec rok
uvnitř eu
miliarda kč
rok kdy
řecko itálie
člen eu
říci ln
čtk
rámec eu
kvóta přerozdělování
mutace metro
rada eu
životní prostředí
ministr zahraniční
eu nato
eu země
zahraničí v

Unsupervised

- Wordfish
  - Slapin & Proksch 2008
  - Scaling algorithm
  - Logic of the process
    - Assume there is a prior distribution of words
    - Let computer find patterns in the distribution of words from documents
    - Estimate importance of individual words
    - Scale documents based on the importance of terms
  - High reliability, but questionable validity
  - Only single dimension, single issue requirement
  - Word meanings have to be stable over time

Unsupervised – Wordfish

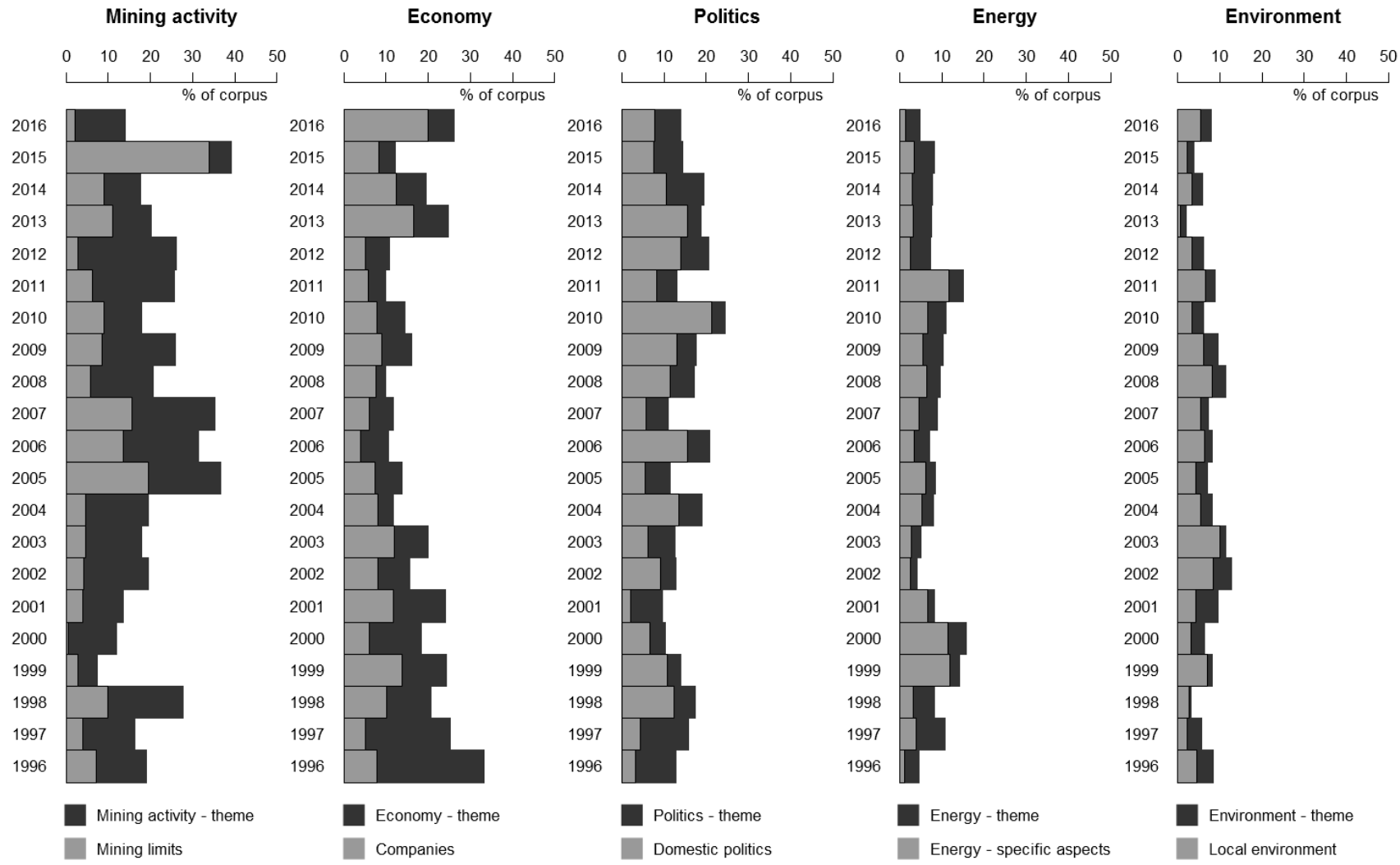




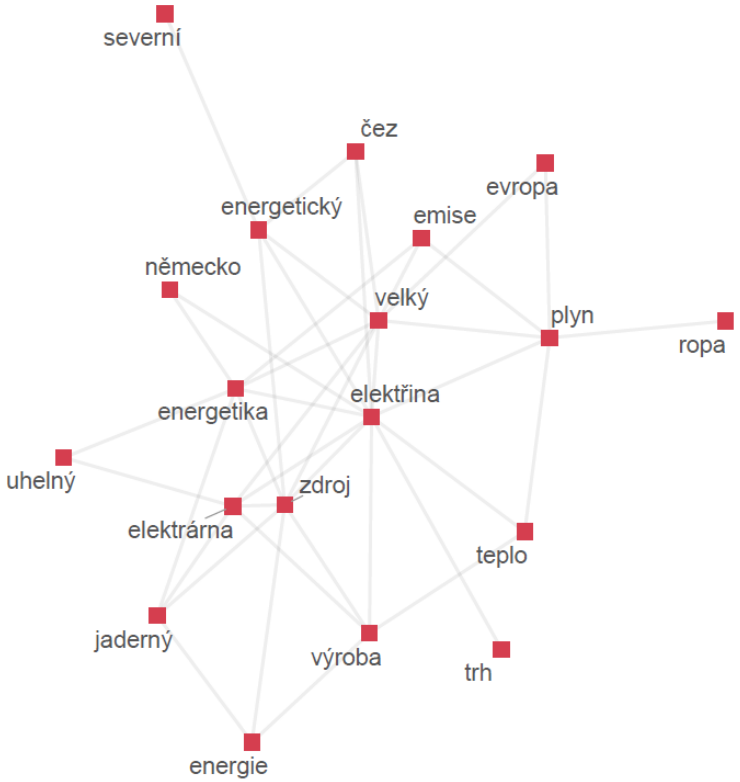
Unsupervised

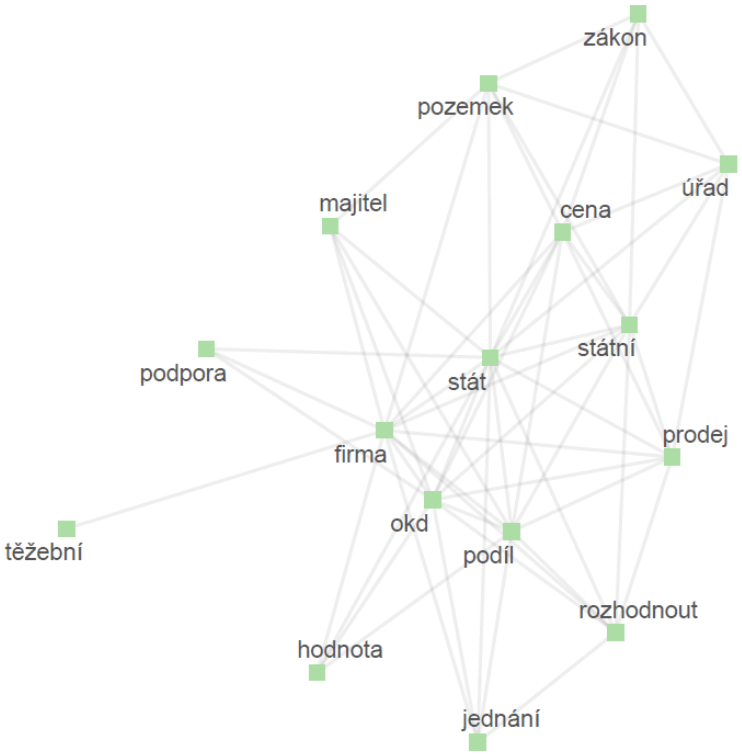
- Topic modeling
  - Blei, Ng and Jordan 2003
  - Clustering algorithm
  - Logic of the process
    - Assume there is a prior distribution of topics over documents and words over topics
    - Let computer find patterns in the distribution of words from documents
    - Estimate the distributions
  - Prior assumption of known number of topics
  - Questionable validity
  - Computer intensive
  - Sensitivities of model setup

Unsupervised – topic modeling



Unsupervised – co-occurrence net





Discourse network analyzer

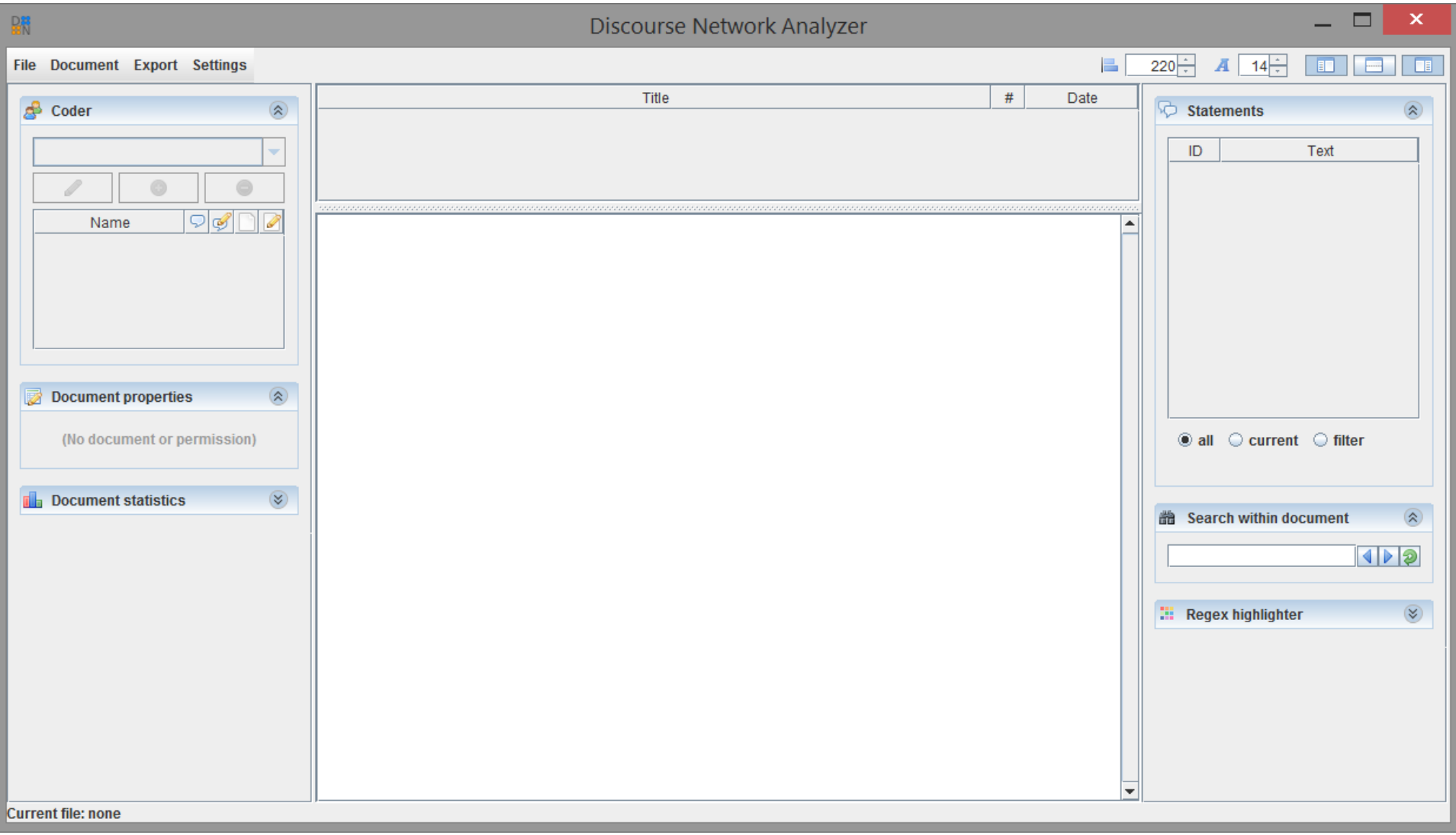
https://github.com/leifeld/dna

Getting started

- Benefits
  - Suitable for any coding task
  - Open source and free of charge for academic purposes
  - Multicoder support
  - Open to variable construction
  - R integration
- Installation
  - Requires working Java JDK
    - https://www.java.com/en/download/manual.jsp
  - See the software manual for details
    - https://github.com/leifeld/dna/releases/download/v2.0-beta.21/dna-manual.pdf

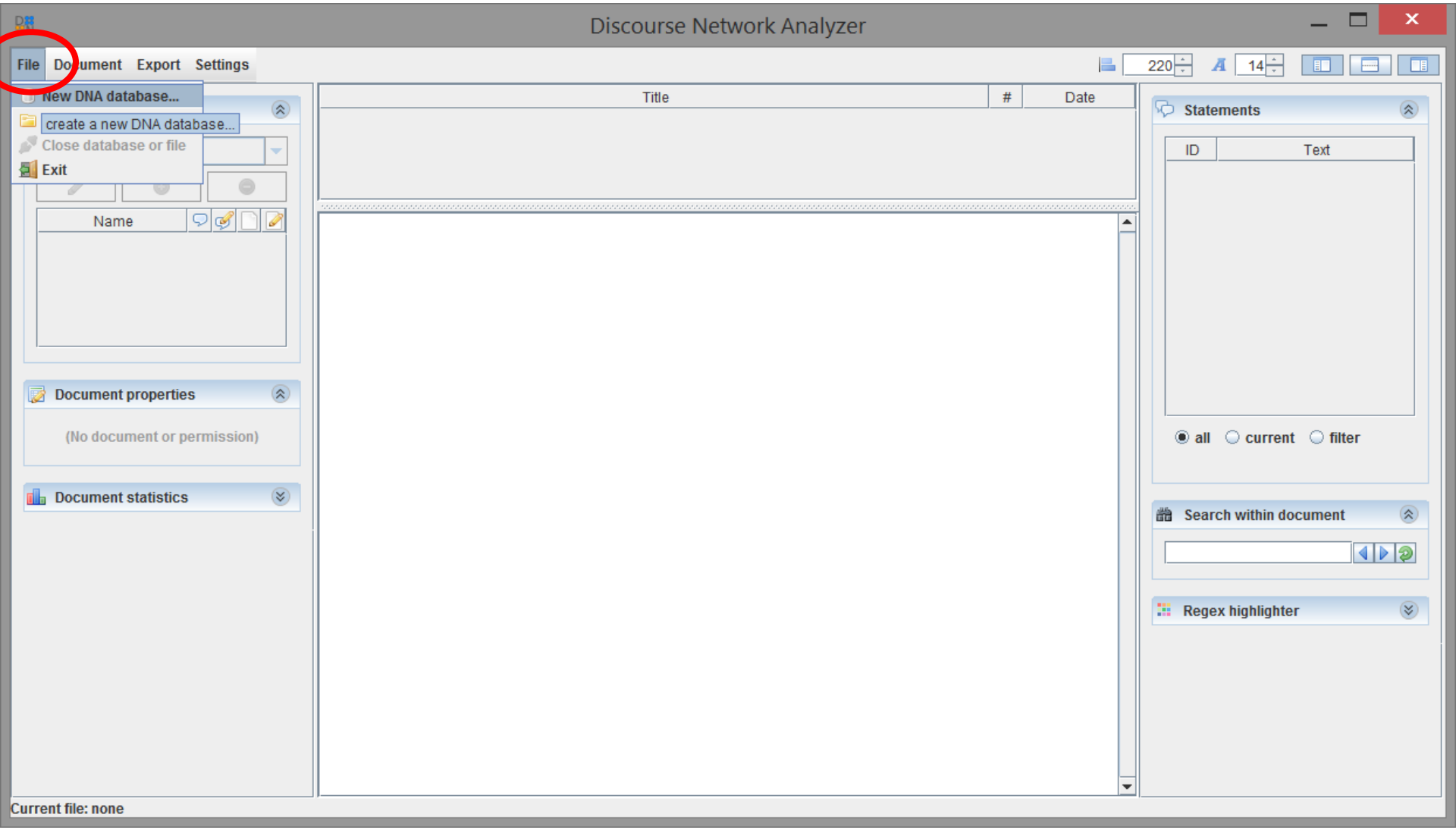
Before we start

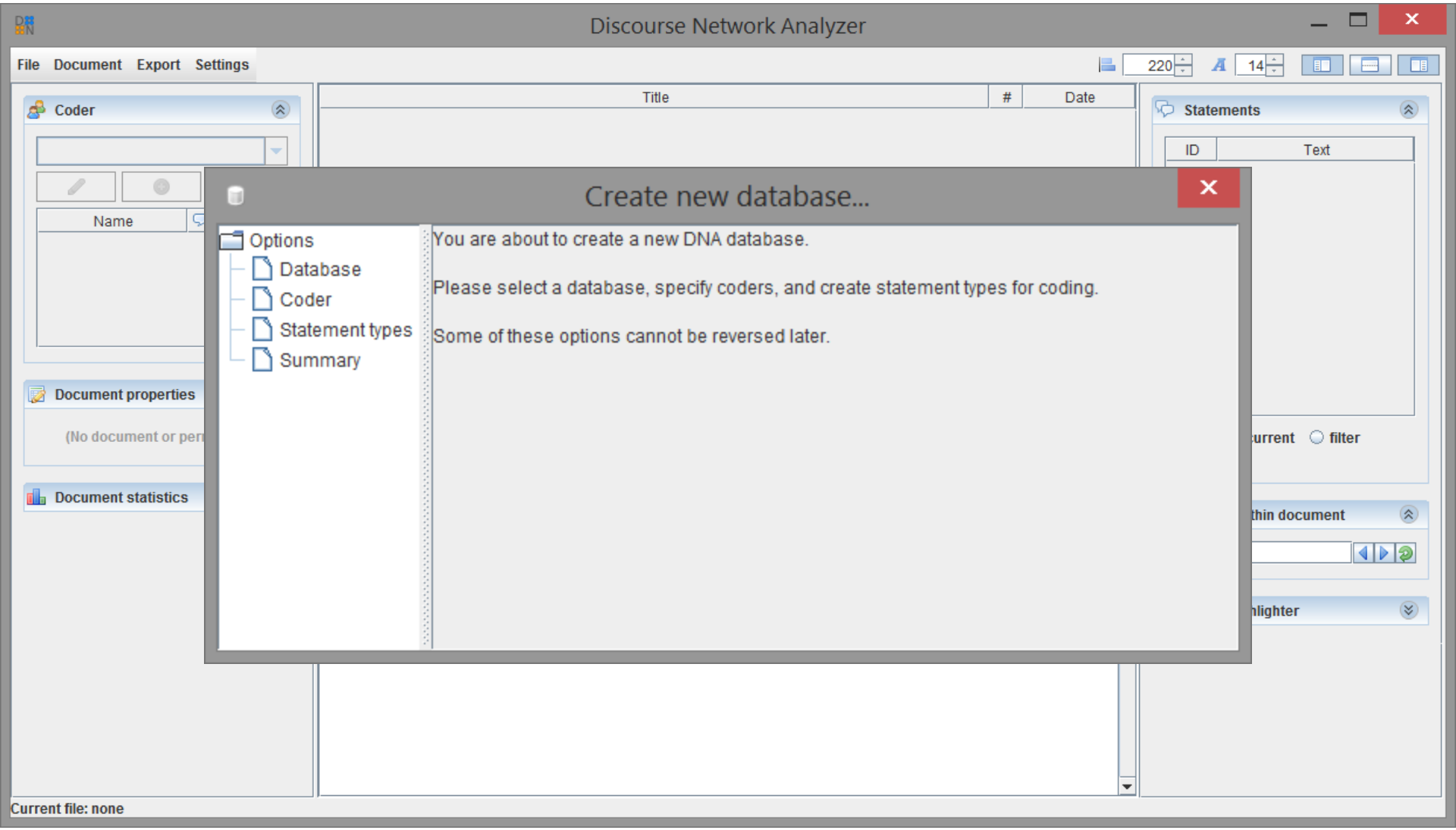
- Download text files to your computer
  - Package “text_analysis_1.zip”
- Unpack all files into an easily-accessible folder
- Open the folder “text_analysis_quali”
- Try to run the “dna-2.0-beta22.jar” program
  - If it runs, Java is installed on your computer
  - If not, you should install Java first (see the previous slide)

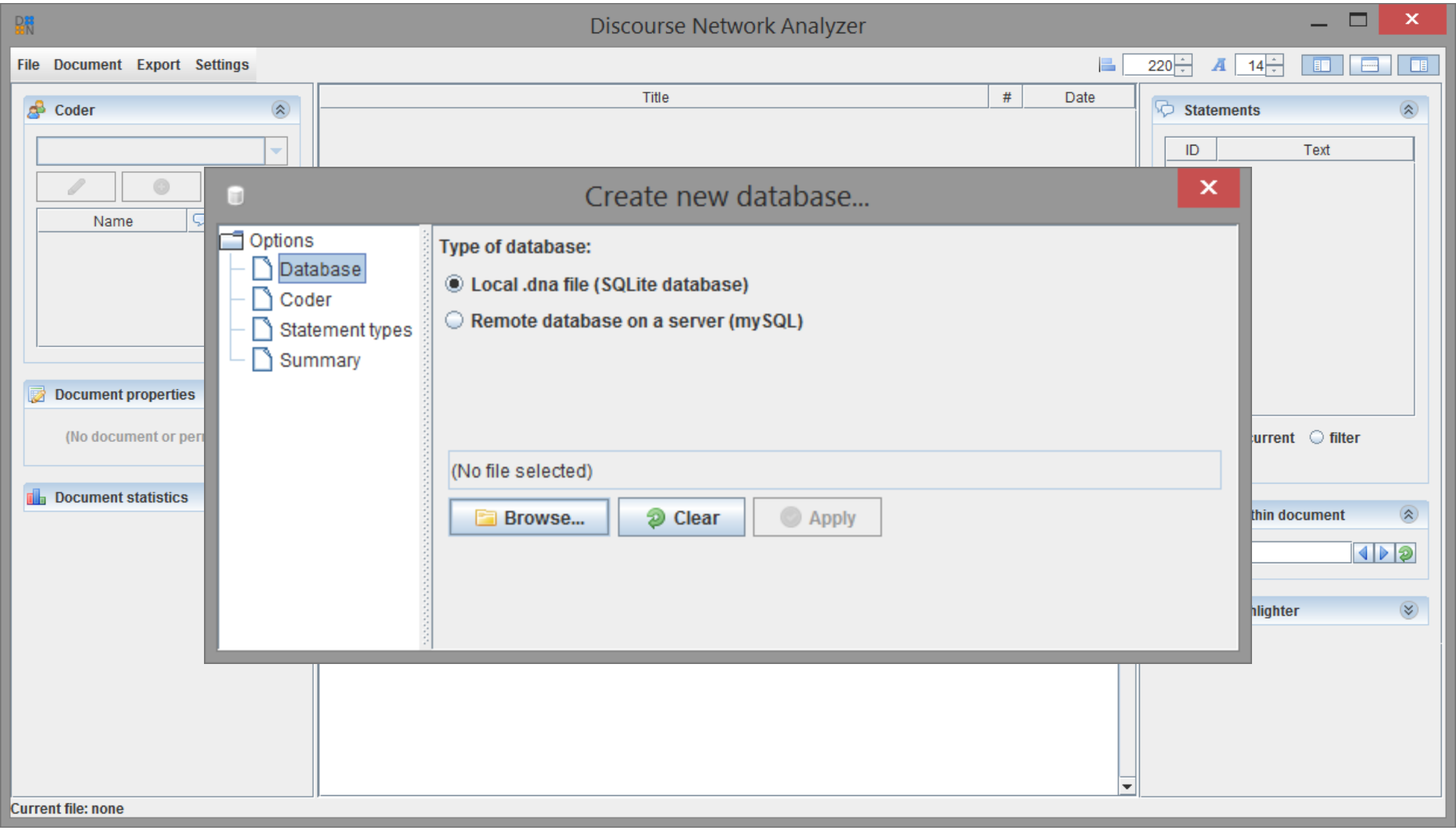

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Name
Document properties
(No document or permission)
Document statistics
Title
#
Date
Statements
ID
Text
all
current
filter
Search within document
Regex highlighter
Current file: none

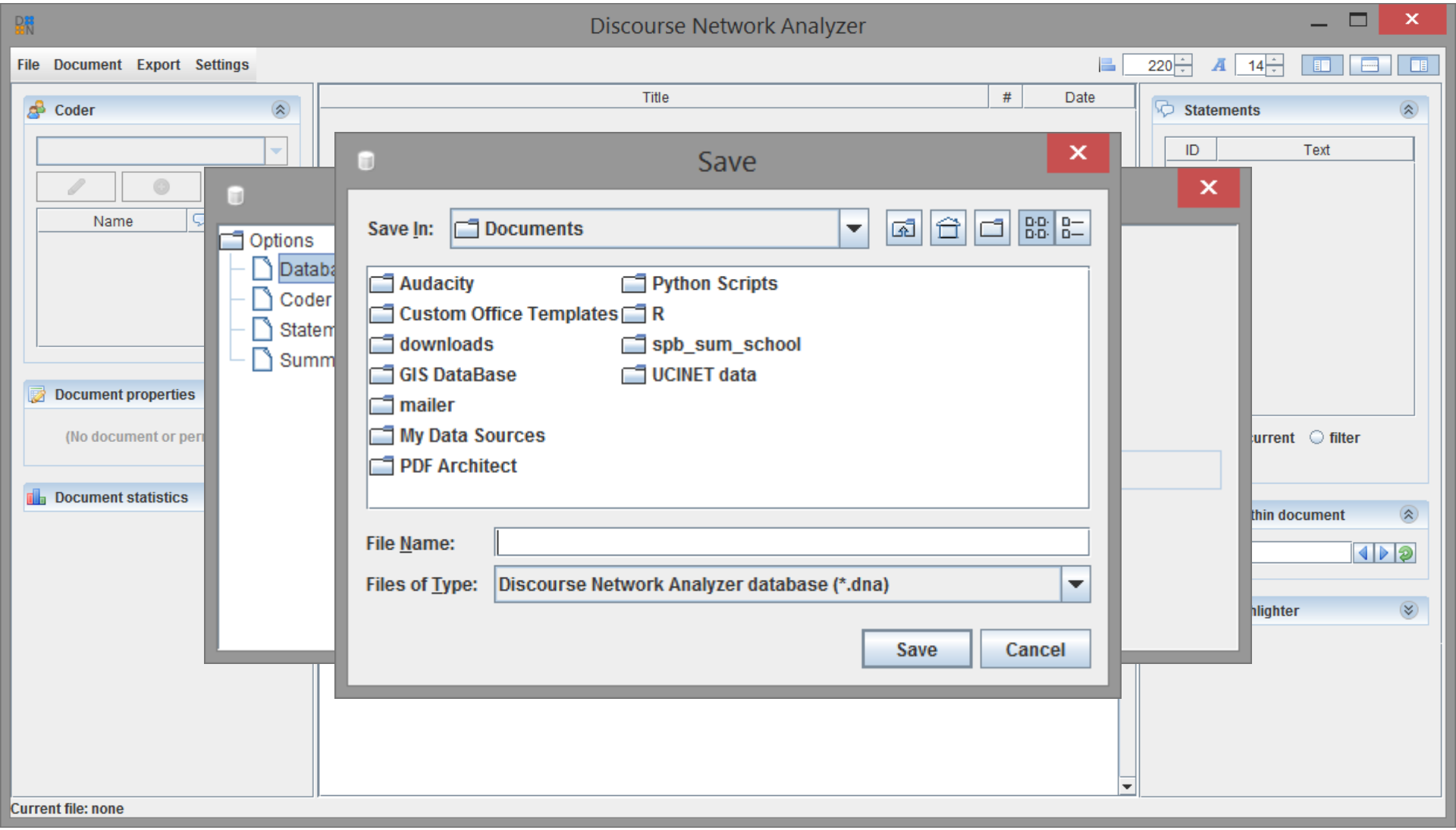
Database

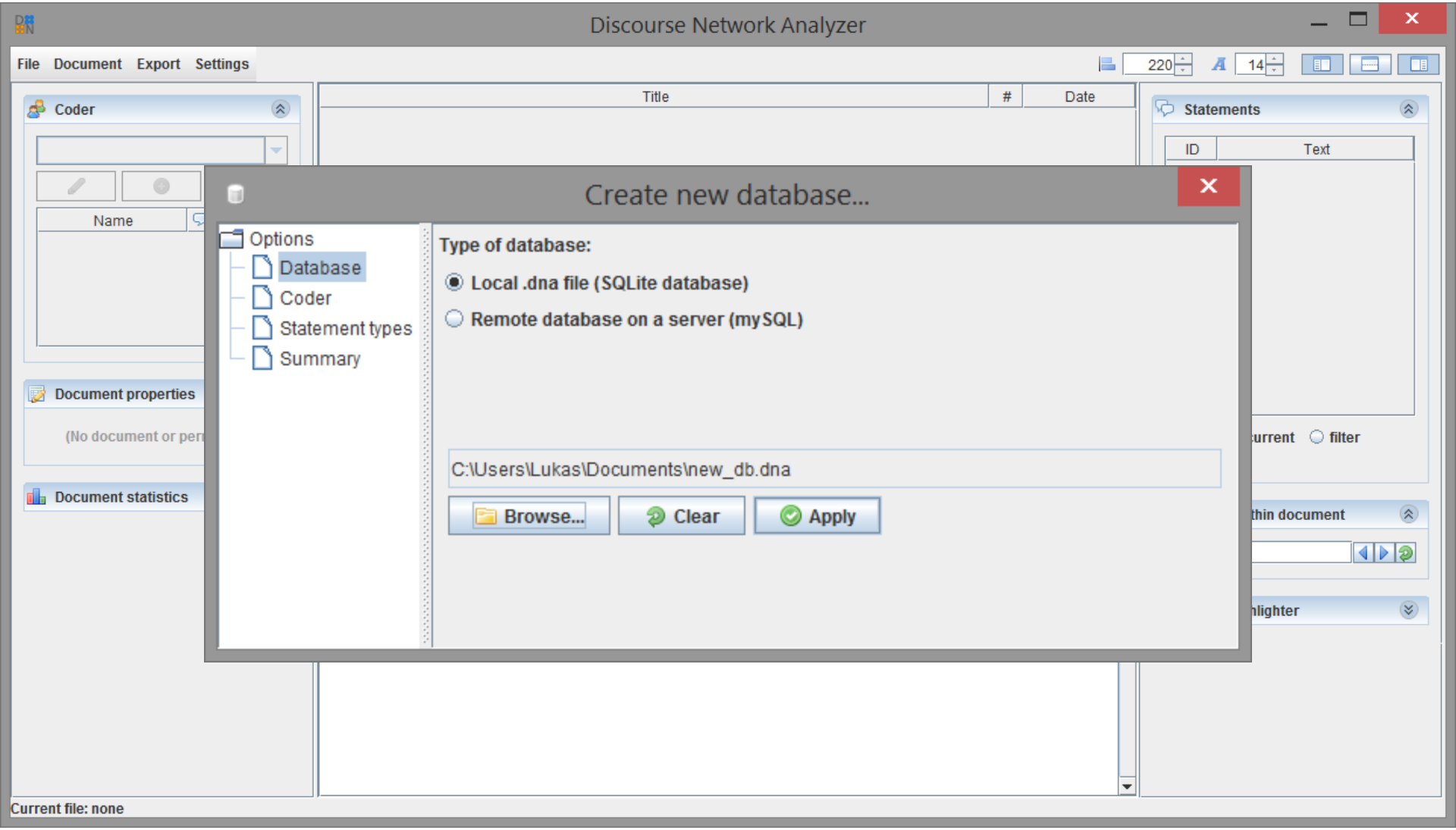
- The format DNA uses is DNA database
  - Identification of number of coders and their names
  - Identification of code categories and variables
- Basic properties of the database can’t be changed later

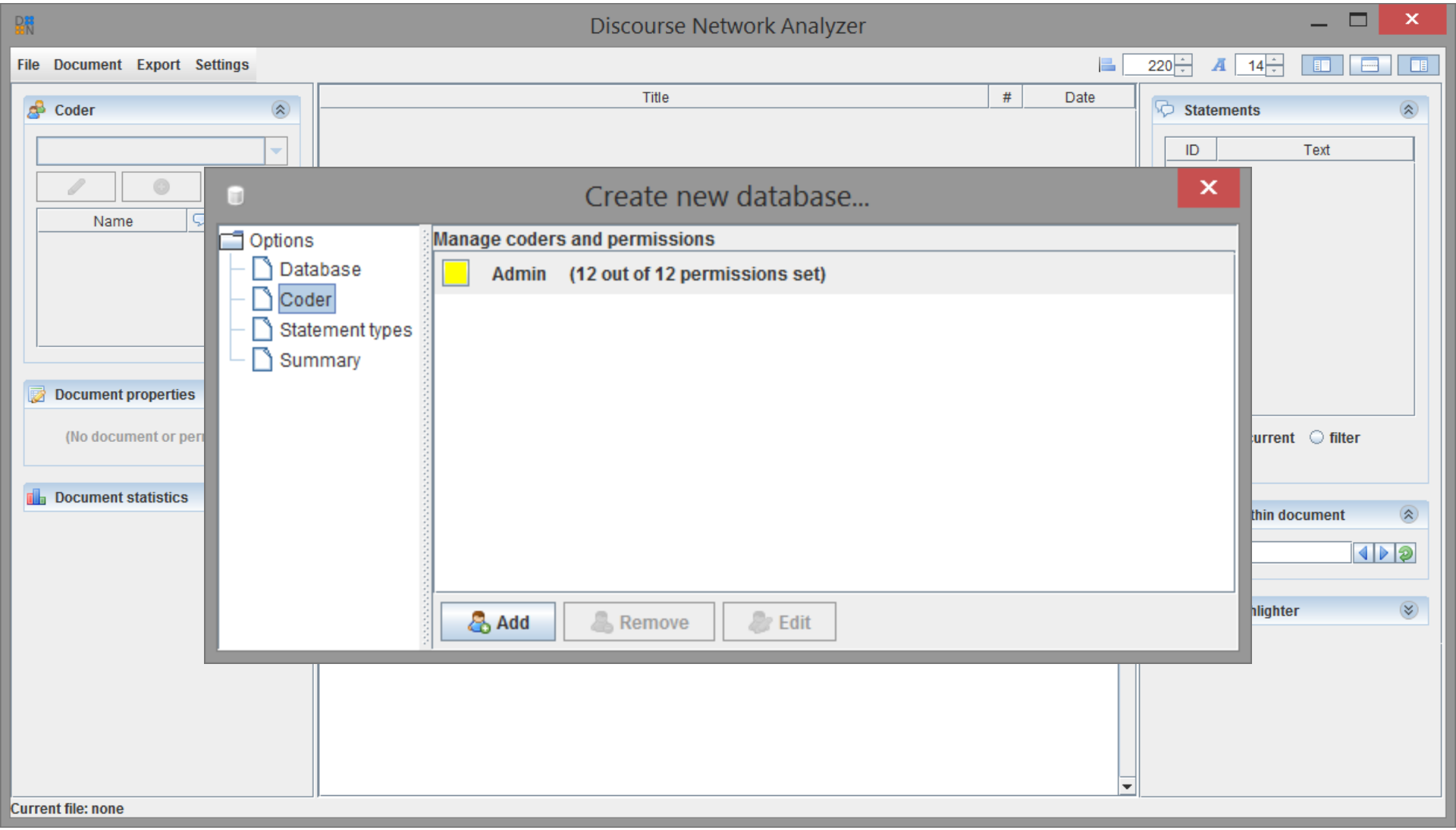

Discourse Network Analyzer
File
Document
Export
Settings
New DNA database...
create a new DNA database...
Close database or file
Exit
220
14
Name
Title
#
Date
Statements
ID
Text
all
current
filter
Document properties
(No document or permission)
Document statistics
Search within document
Regex highlighter
Current file: none

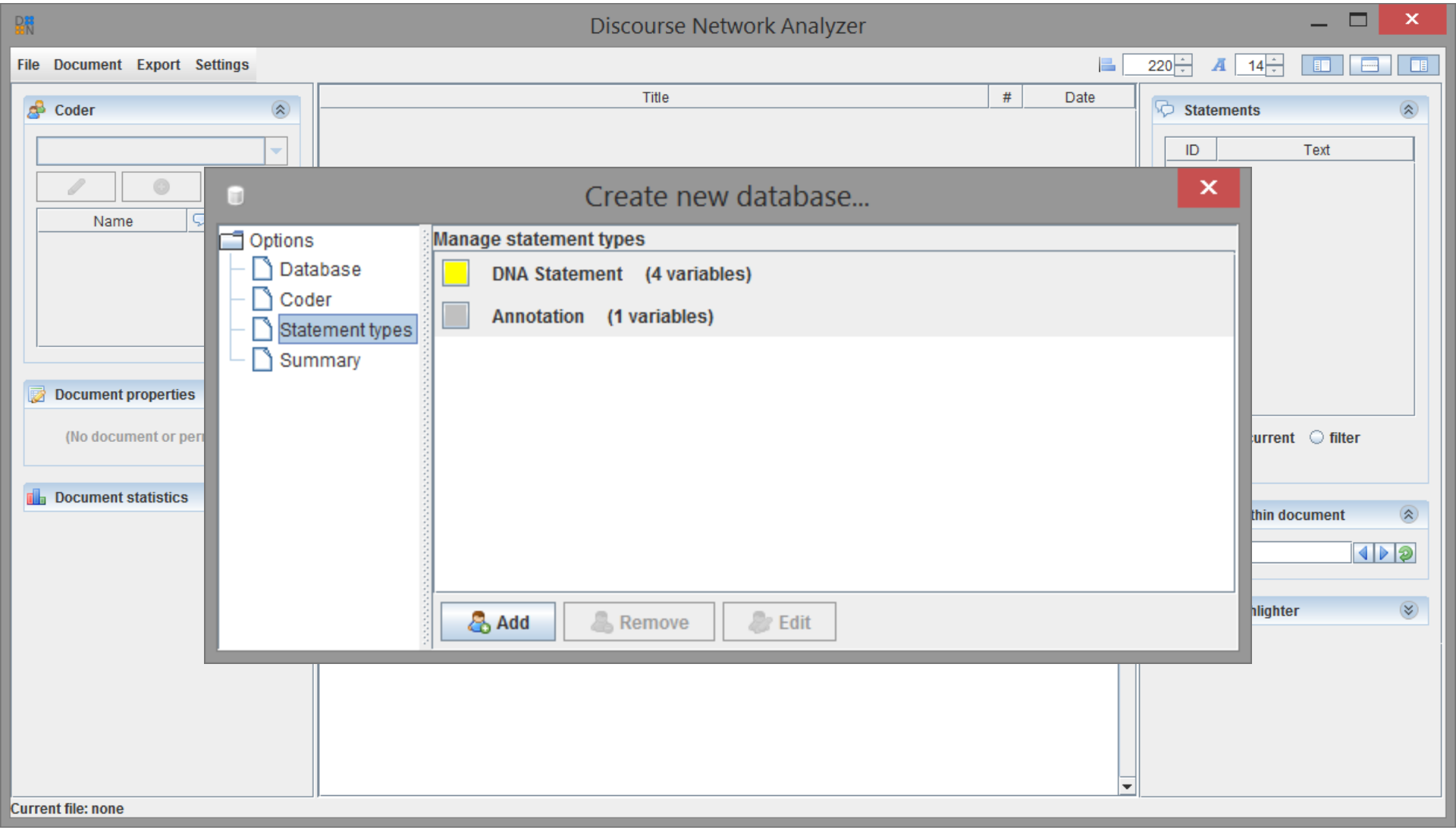

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Name
Title
#
Date
Statements
ID
Text
Document properties
(No document or per
Document statistics
current
filter
thin document
hlighter
Create new database...
Options
Database
Coder
Statement types
Summary
You are about to create a new DNA database.
Please select a database, specify coders, and create statement types for coding.
Some of these options cannot be reversed later.
Current file: none

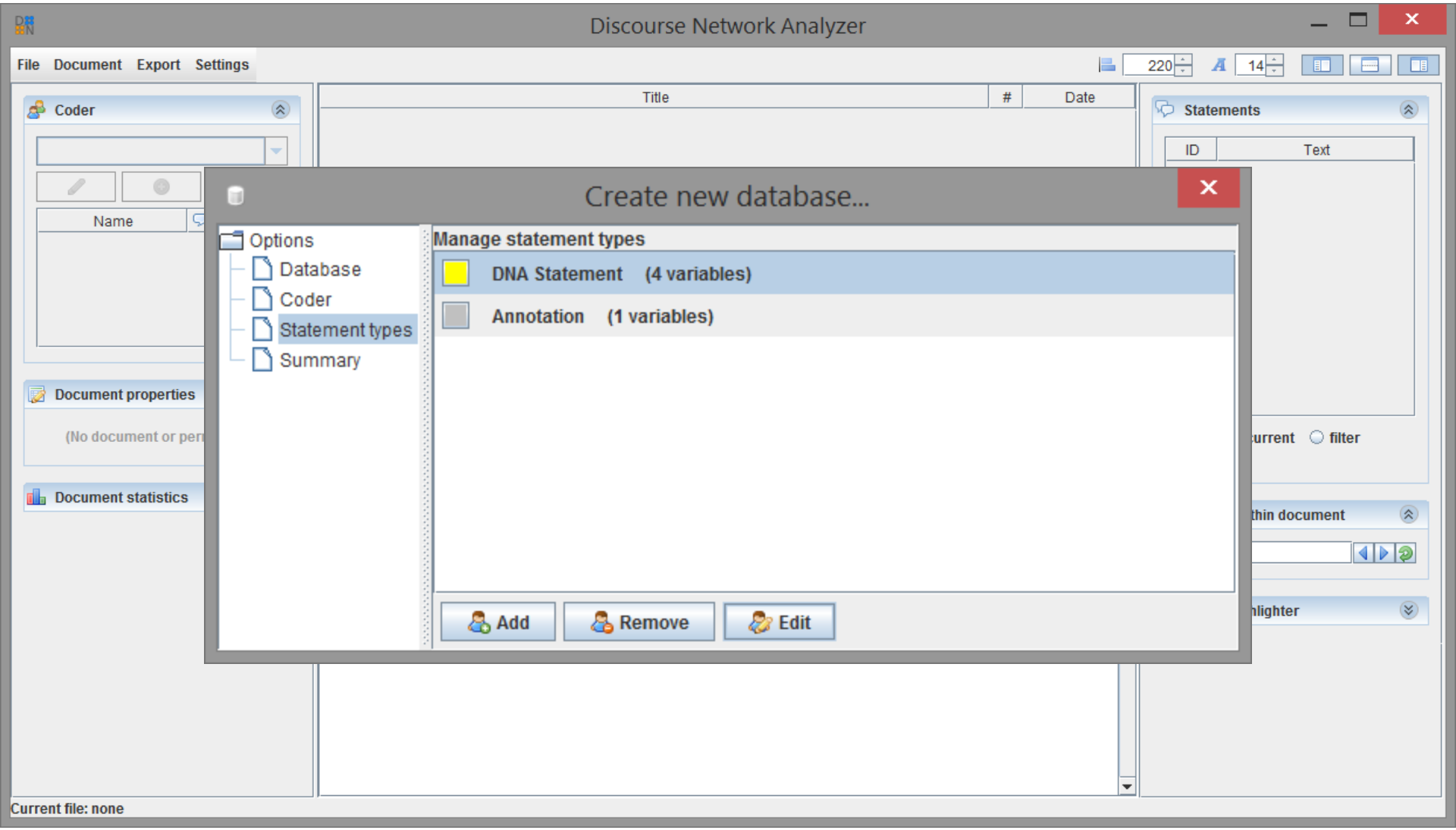

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Name
Title
#
Date
Statements
ID
Text
Document properties
(No document or per
Document statistics
current
filter
thin document
hlighter
Current file: none
Create new database...
Options
Database
Coder
Statement types
Summary
Type of database:
Local .dna file (SQLite database)
Remote database on a server (mySQL)
(No file selected)
Browse...
Clear
Apply

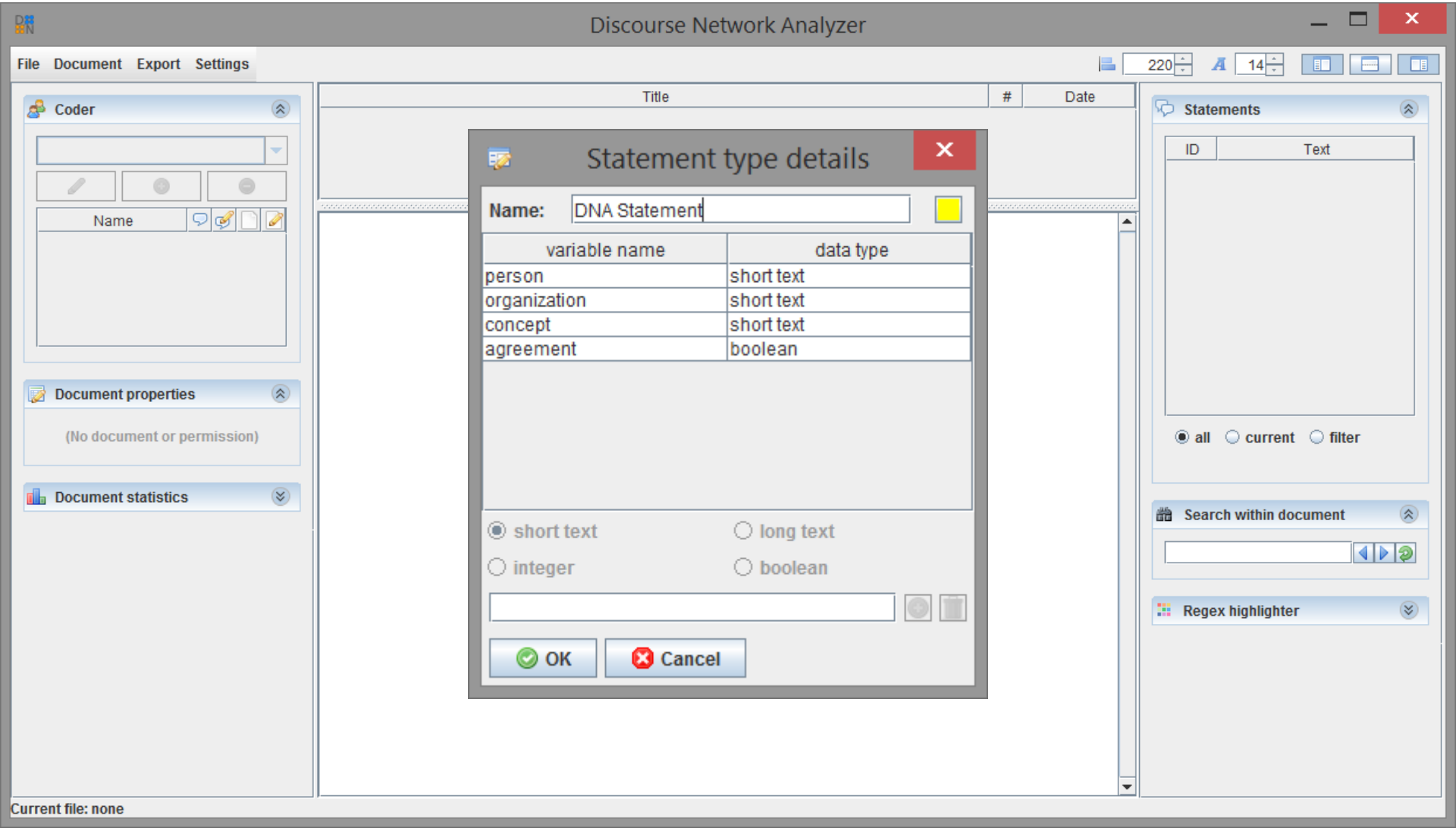

Discourse Network Analyzer
File
Document
Export
Settings
Coder
Name
Document properties
Document statistics
Title
#
Date
Statements
ID
Text
Options
Save
Save In:
Documents
Audacity
Custom Office Templates
downloads
GIS DataBase
mailer
My Data Sources
PDF Architect
Python Scripts
R
spb_sum_school
UCINET data
File Name:
Files of Type:
Discourse Network Analyzer database (*.dna)
Save
Cancel
Current file: none

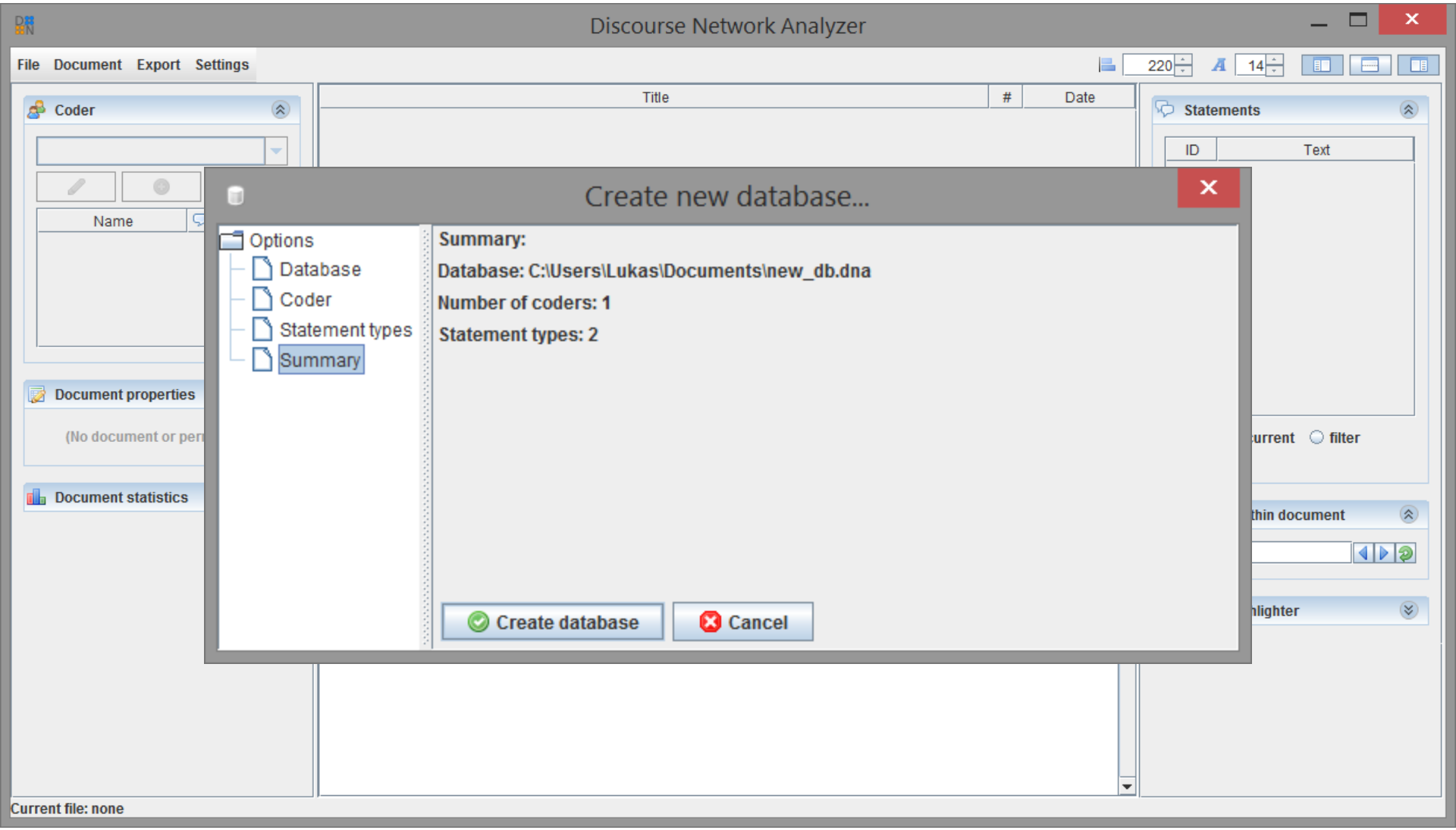

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Name
Title
#
Date
Statements
ID
Text
Document properties
(No document or per
Document statistics
current
filter
thin document
hlighter
Create new database...
Options
Database
Coder
Statement types
Summary
Type of database:
Local .dna file (SQLite database)
Remote database on a server (mySQL)
C:\Users\Lukas\Documents\new_db.dna
Browse...
Clear
Apply
Current file: none


Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Name
Title
#
Date
Statements
ID
Text
Document properties
(No document or per
Document statistics
current
filter
thin document
hlighter
Create new database...
Options
Database
Coder
Statement types
Summary
Manage coders and permissions
Admin (12 out of 12 permissions set)
Add
Remove
Edit
Current file: none


Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Name
Title
#
Date
Statements
ID
Text
Document properties
(No document or per
Document statistics
current
filter
thin document
hlighter
Create new database...
Options
Database
Coder
Statement types
Summary
Manage statement types
DNA Statement (4 variables)
Annotation (1 variables)
Add
Remove
Edit
Current file: none

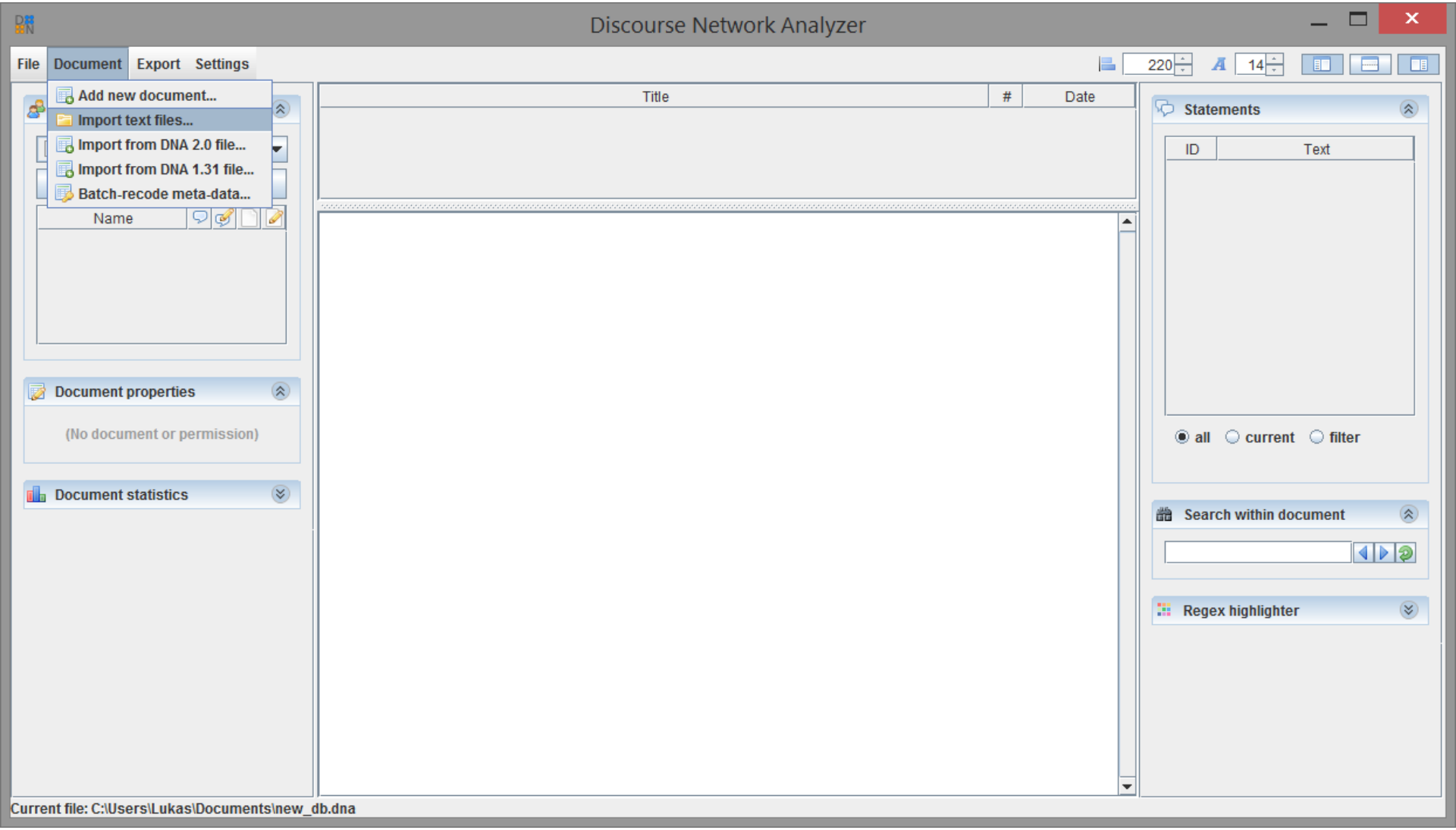

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Name
Title
#
Date
Statements
ID
Text
Document properties
(No document or per
Document statistics
current
filter
thin document
hlighter
Create new database...
Options
Database
Coder
Statement types
Summary
Manage statement types
DNA Statement (4 variables)
Annotation (1 variables)
Add
Remove
Edit
Current file: none

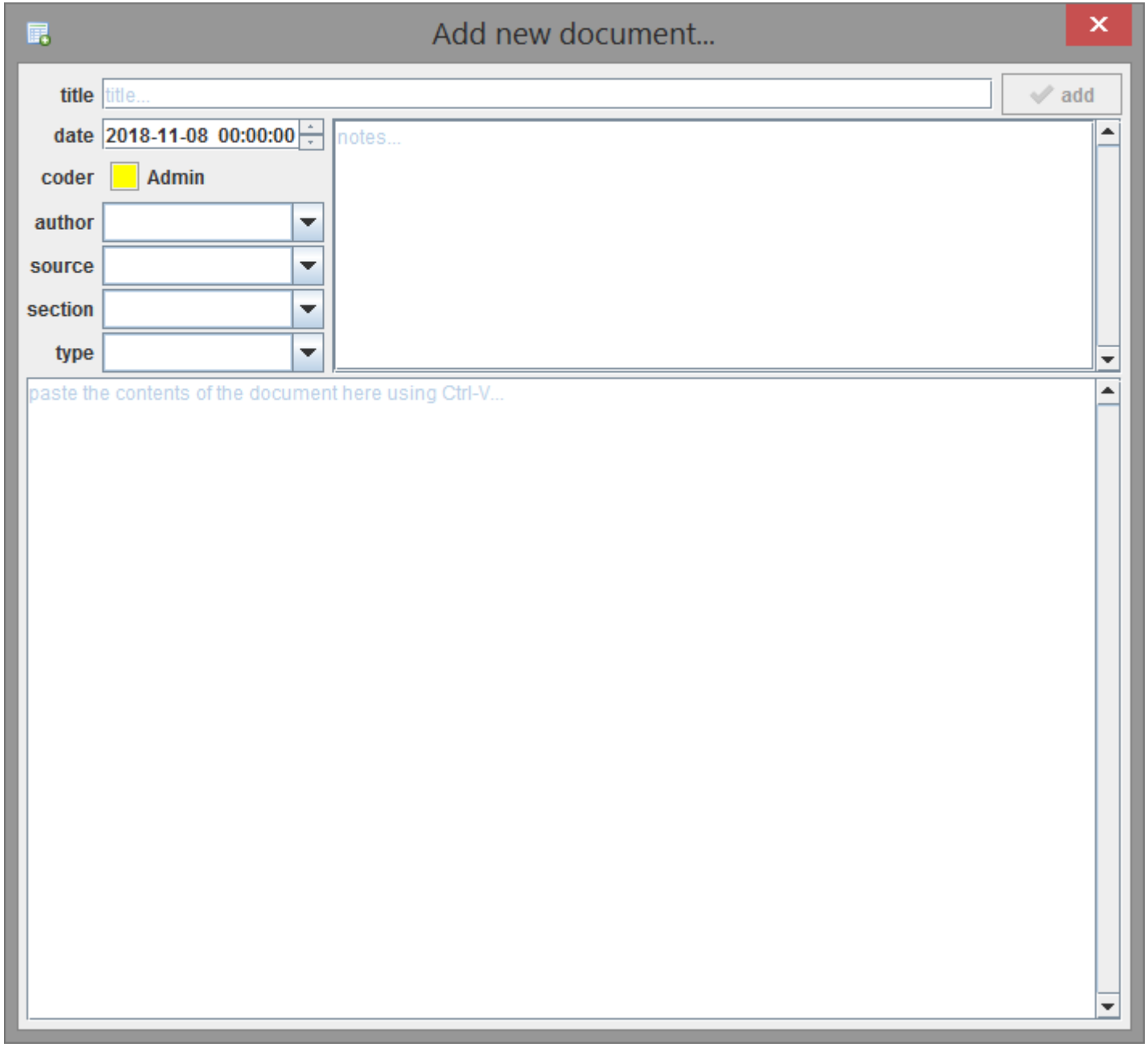

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Name
Document properties
(No document or permission)
Document statistics
Title
#
Date
Statement type details
Name:
DNA Statement
variable name
data type
person
short text
organization
short text
concept
short text
agreement
boolean
short text
long text
integer
boolean
OK
Cancel
Statements
ID
Text
all
current
filter
Search within document
Regex highlighter
Current file: none

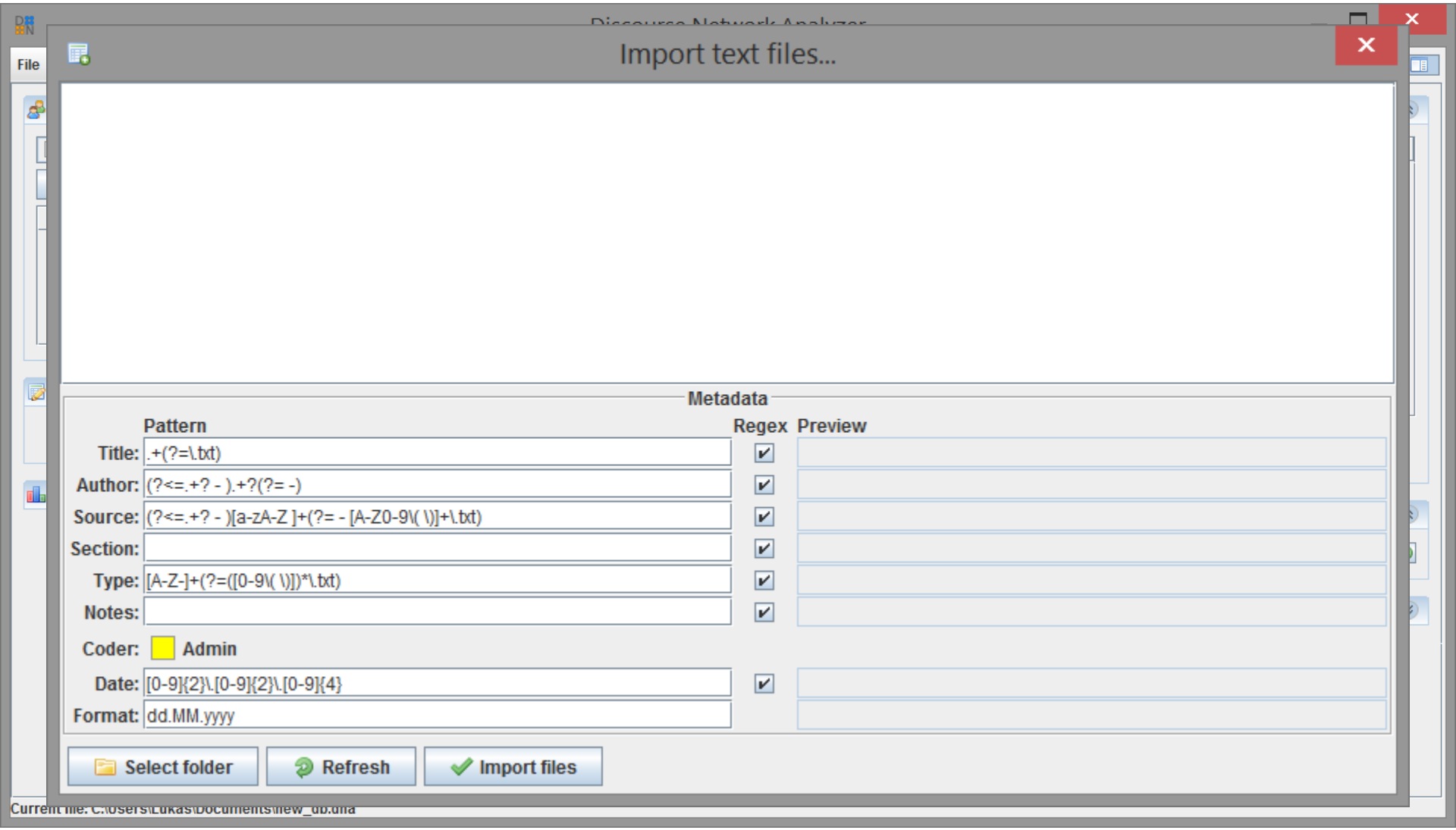

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Name
Title
#
Date
Statements
ID
Text
Document properties
(No document or per
Document statistics
current
filter
thin document
hlighter
Create new database...
Options
Database
Coder
Statement types
Summary
Summary:
Database: C:\Users\Lukas\Documents\new_db.dna
Number of coders: 1
Statement types: 2
Create database
Cancel
Current file: none

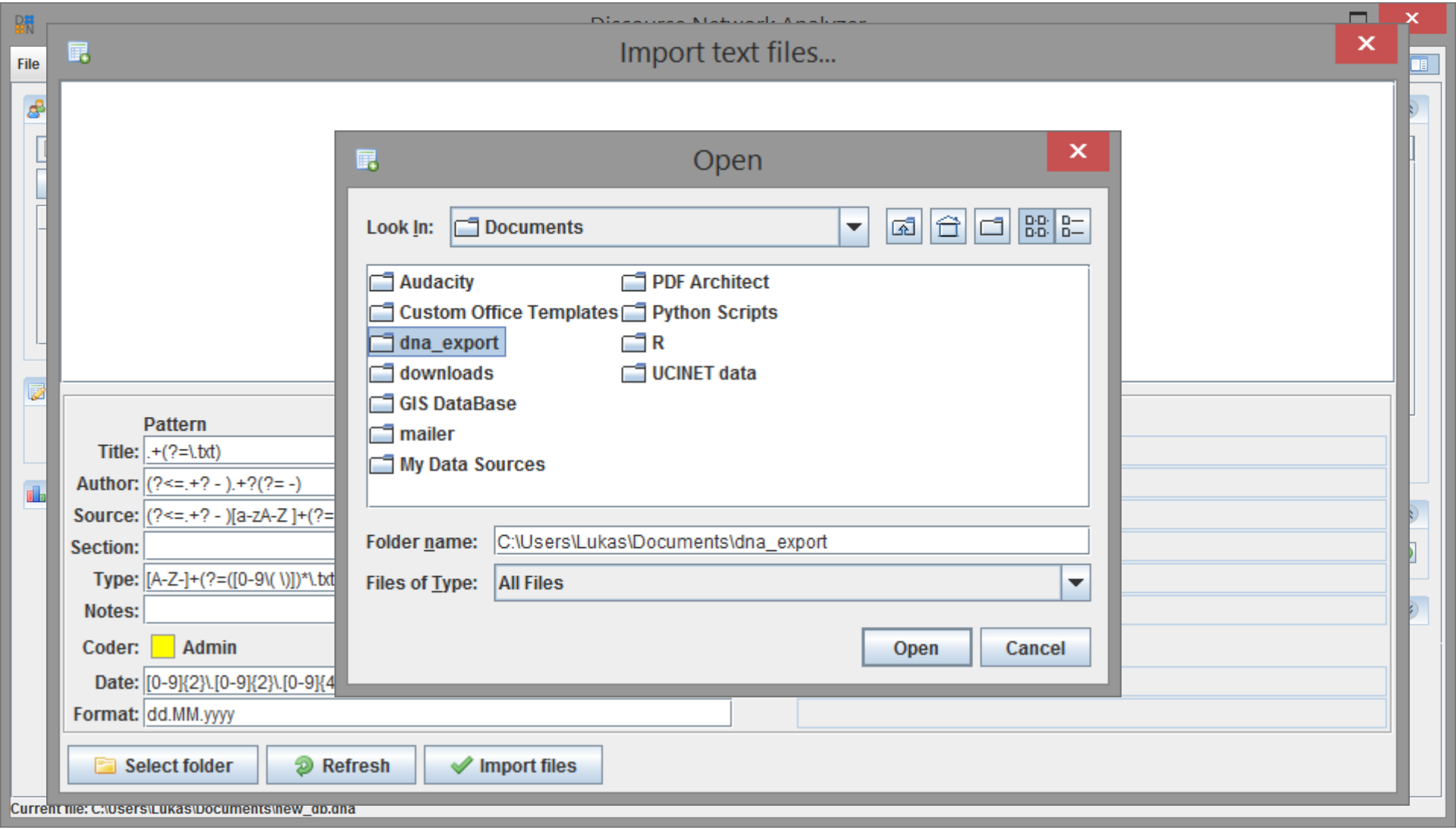
Data import

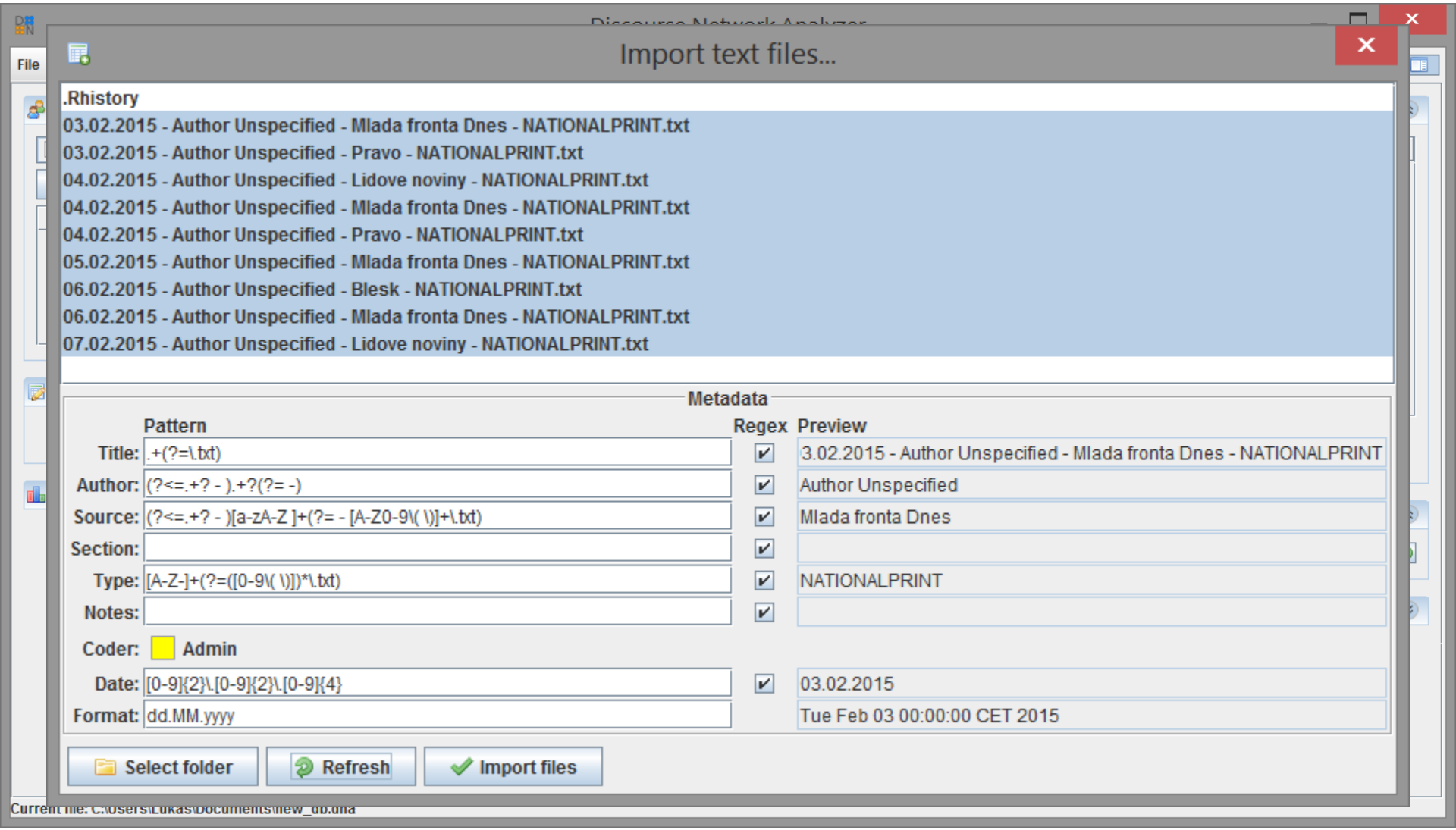
- Manual import of data – text by text
- Folder import
  - Place all documents in one folder (text format)
  - Identify folder
  - Documents get loaded into the software
- Automatic construction of metadata
  - Proper document name results into automatic metadata construction
  - dd.mm.yyyy – Name Surname – Publication name – TEXTTYPE.txt

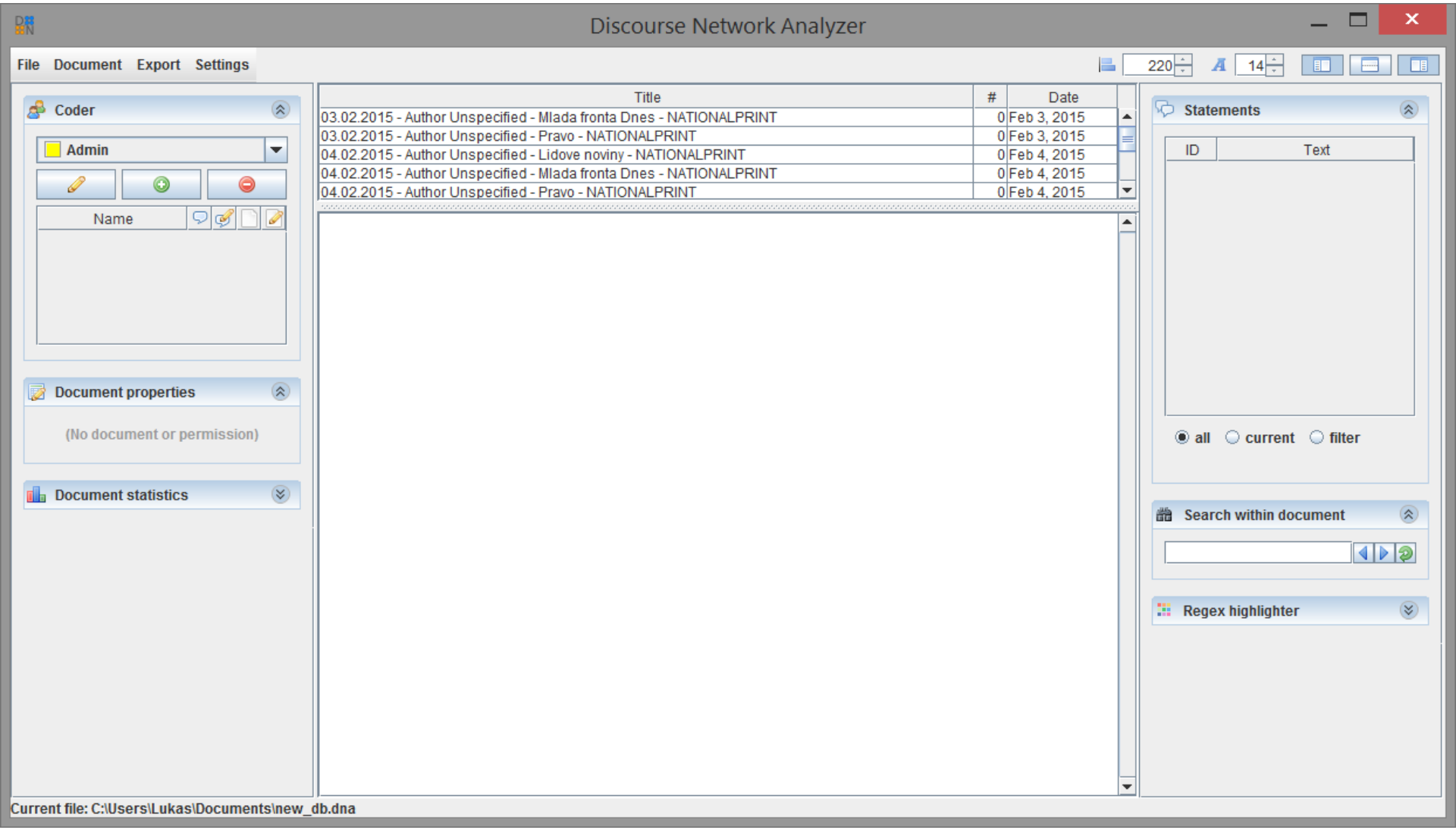

Discourse Network Analyzer
File
Document
Export
Settings
Add new document...
Import text files...
Import from DNA 2.0 file...
Import from DNA 1.31 file...
Batch-recode meta-data...
220
14
Name
Title
#
Date
Statements
ID
Text
all
current
filter
Document properties
(No document or permission)
Document statistics
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna

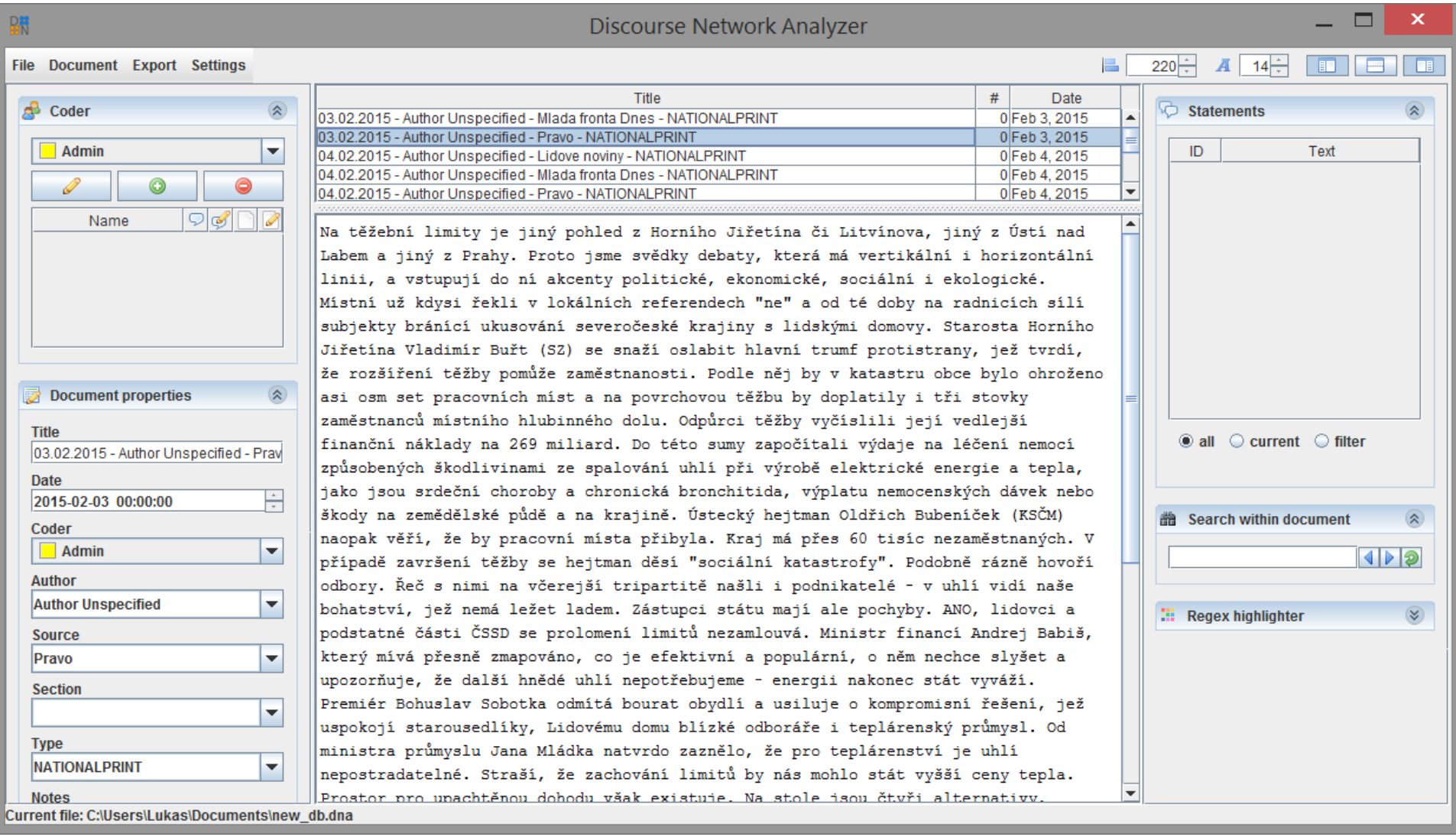

Add new document...
title
title...
add
date
2018-11-08 00:00:00
notes...
coder
Admin
author
source
section
type
paste the contents of the document here using Ctrl-V...

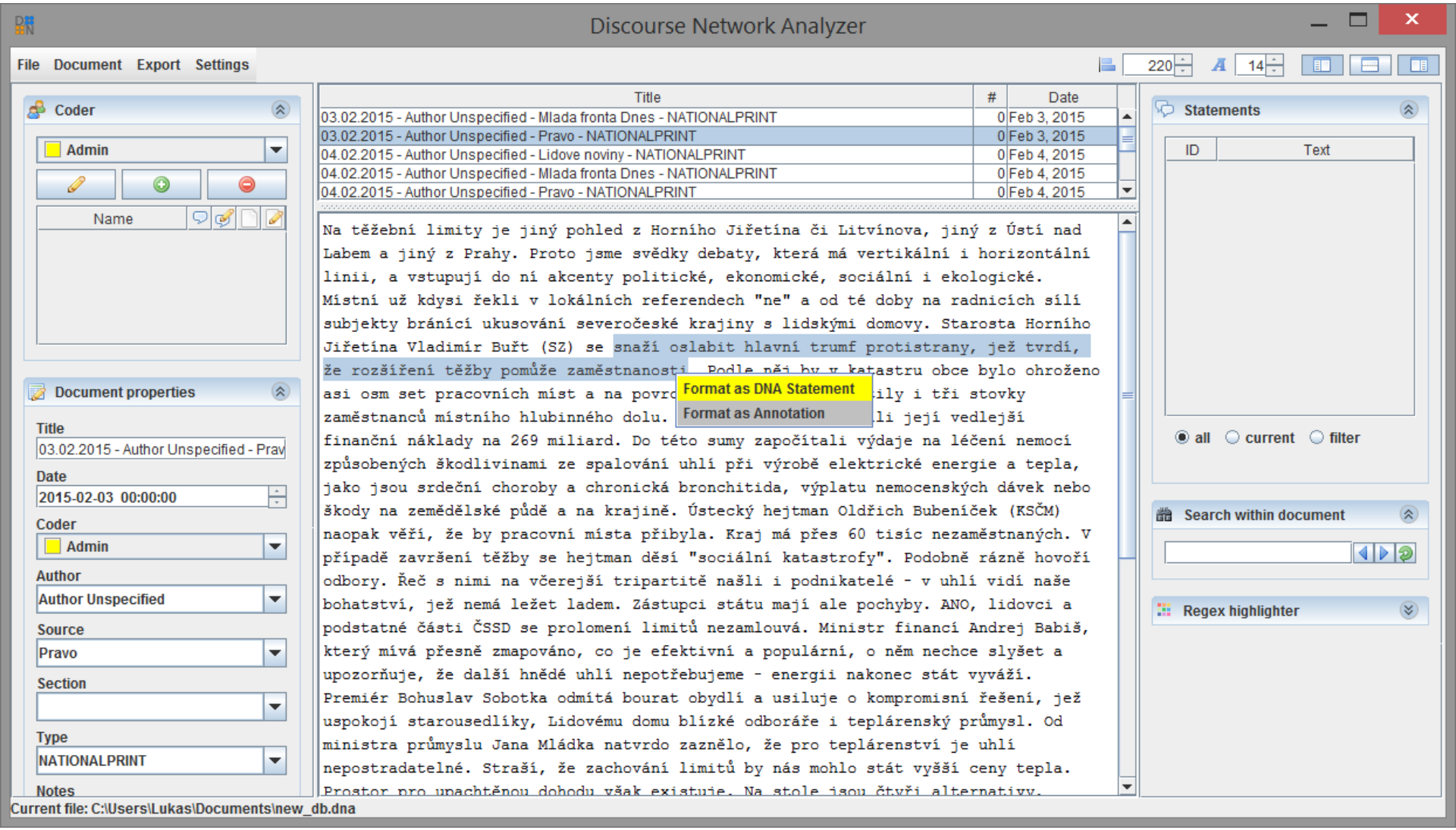

Import text files...
Metadata
Pattern
Regex
Preview
Title: .+(?=\.txt)
Author: (?<=.+? -).+?(?= -)
Source: (?<=.+? -)[a-zA-Z]+(?= - [A-Z0-9\(\)]+\.txt)
Section:
Type: [A-Z-]+(?=([0-9\(\)])*\.txt)
Notes:
Coder: Admin
Date: [0-9]{2}\.[0-9]{2}\.[0-9]{4}
Format: dd.MM.yyyy
Select folder
Refresh
Import files

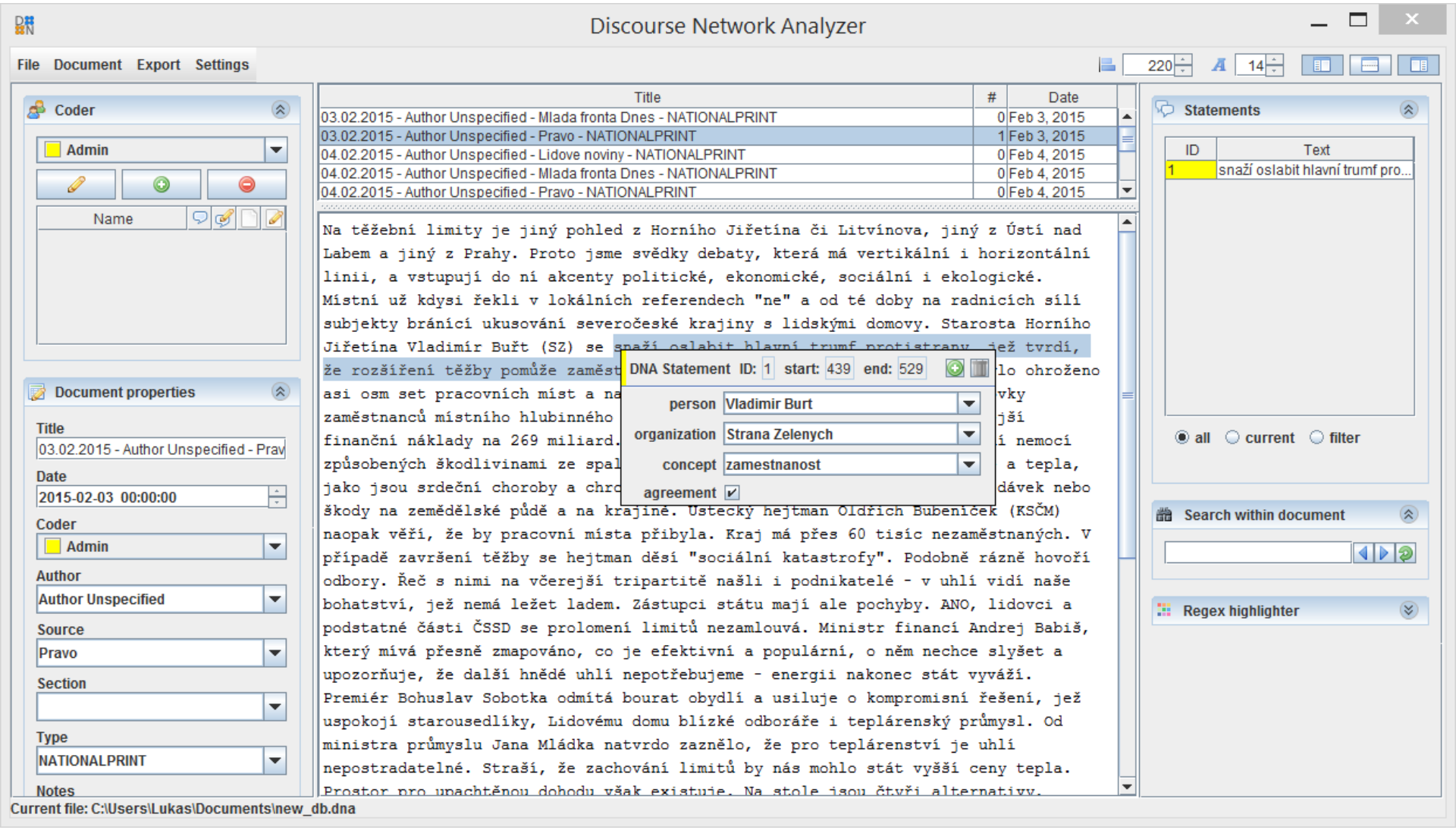

Import text files...
Open
Look In:
Documents
Audacity
Custom Office Templates
dna_export
downloads
GIS DataBase
mailer
My Data Sources
PDF Architect
Python Scripts
R
UCINET data
Folder name:
C:\Users\Lukas\Documents\dna_export
Files of Type:
All Files
Open
Cancel
Pattern
Title:
.+(?=\.txt)
Author:
(?<=.+? -).+?(?= -)
Source:
(?<=.+? -)[a-zA-Z]+(?=
Section:
Type:
[A-Z-]+(?=([0-9\(\)])*\.txt
Notes:
Coder:
Admin
Date:
[0-9]{2}\.[0-9]{2}\.[0-9]{4
Format:
dd.MM.yyyy
Select folder
Refresh
Import files
File
Current file: C:\Users\Lukas\Documents\new_db.dna

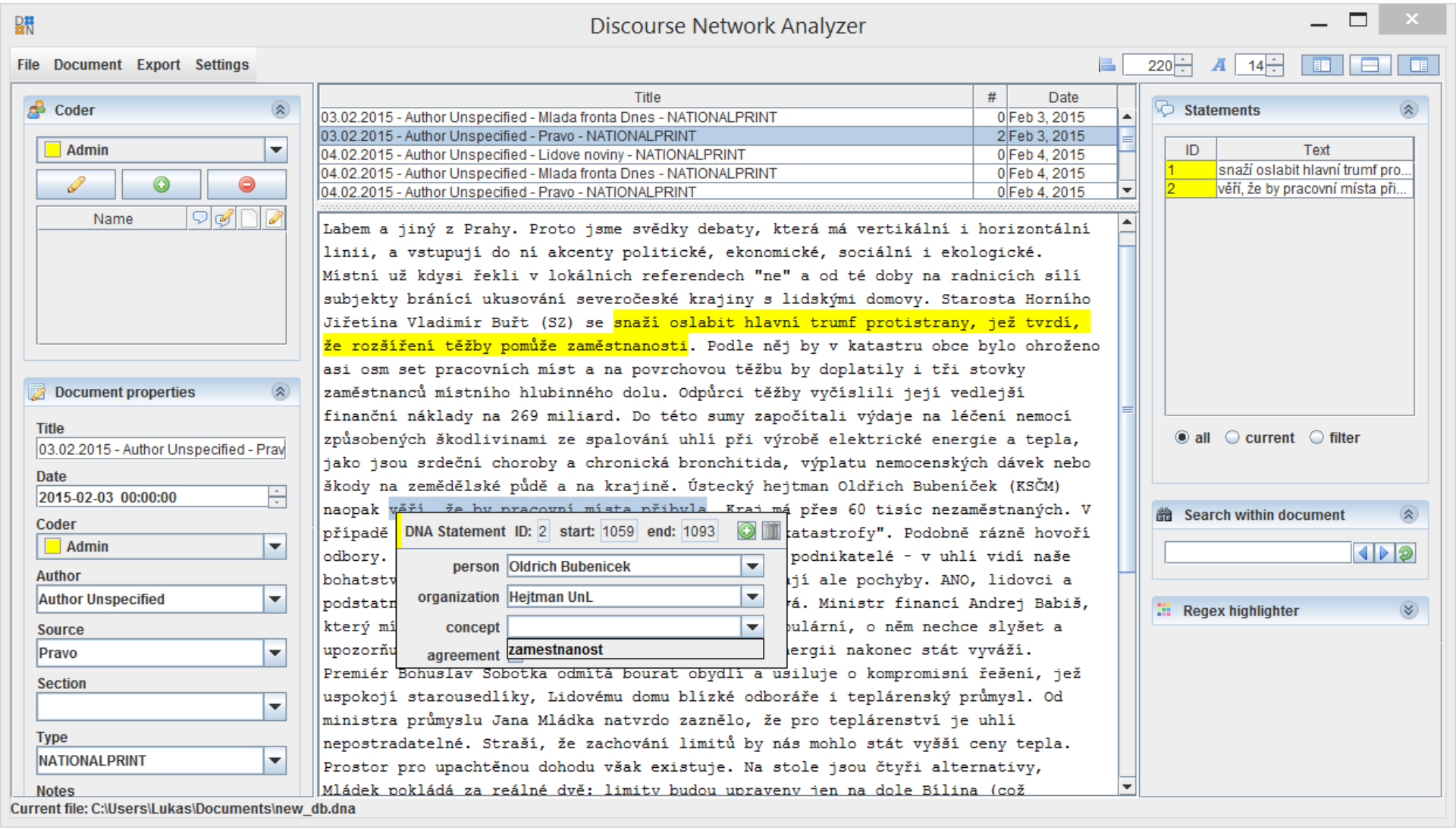

Import text files...
.Rhistory
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT.txt
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT.txt
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT.txt
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT.txt
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT.txt
05.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT.txt
06.02.2015 - Author Unspecified - Blesk - NATIONALPRINT.txt
06.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT.txt
07.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT.txt
Metadata
Pattern
Regex
Preview
Title: .+(?=\.txt)
3.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT
Author: (?<=.+? -).+?(?= -)
Author Unspecified
Source: (?<=.+? -)[a-zA-Z]+(?= - [A-Z0-9\(\)]+\.txt)
Mlada fronta Dnes
Section:
Type: [A-Z-]+(?=([0-9\(\)])*\.txt)
NATIONALPRINT
Notes:
Coder:
Admin
Date: [0-9]{2}\.[0-9]{2}\.[0-9]{4}
03.02.2015
Format: dd.MM.yyyy
Tue Feb 03 00:00:00 CET 2015
Select folder
Refresh
Import files

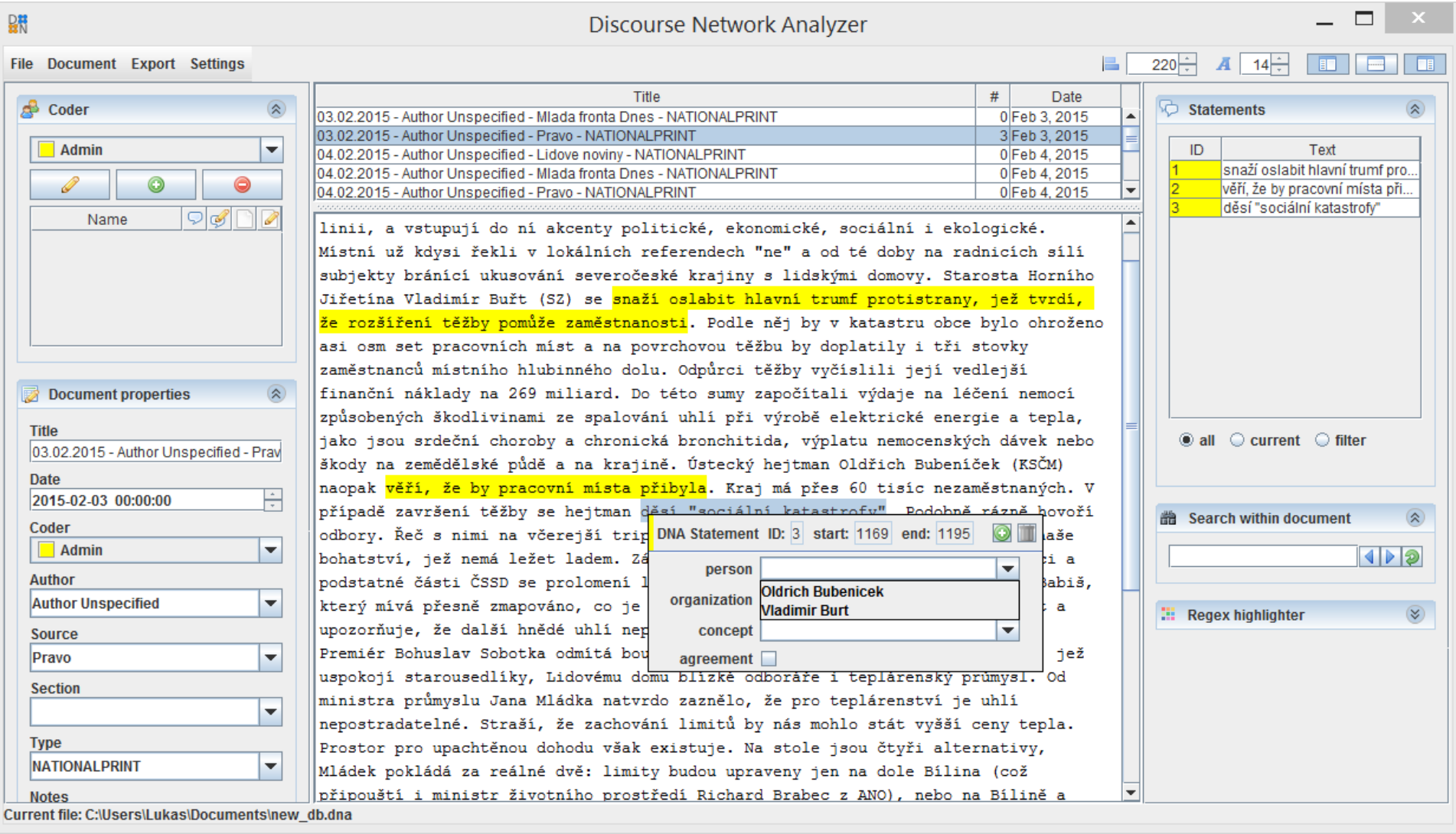

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Admin
Name
Document properties
(No document or permission)
Document statistics
Title
#
Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT
0
Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT
0
Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT
0
Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT
0
Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT
0
Feb 4, 2015
Statements
ID
Text
all
current
filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna

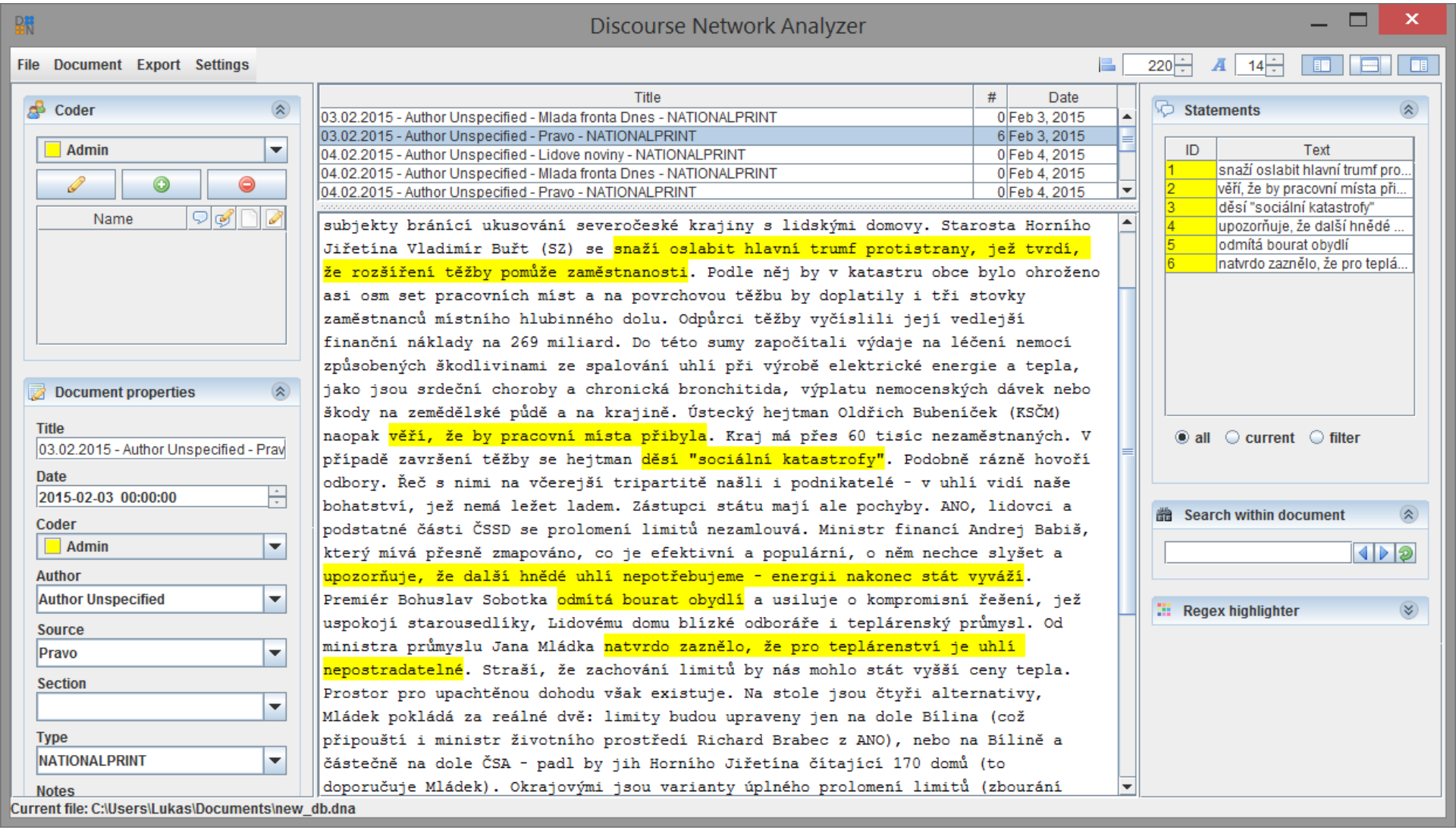

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title
#
Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT
0
Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT
0
Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT
0
Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT
0
Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT
0
Feb 4, 2015
Na těžební limity je jiný pohled z Horního Jiřetína či Litvínova, jiný z Ústí nad Labem a jiný z Prahy. Proto jsme svědky debaty, která má vertikální i horizontální linii, a vstupují do ní akcenty politické, ekonomické, sociální i ekologické. Místní už kdysi řekli v lokálních referendech "ne" a od té doby na radnicích sílí subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního Jiřetína Vladimír Buřt (SZ) se snaží oslabit hlavní trumf protistrany, jež tvrdí, že rozšíření těžby pomůže zaměstnanosti. Podle něj by v katastru obce bylo ohroženo asi osm set pracovních míst a na povrchovou těžbu by doplatily i tři stovky zaměstnanců místního hlubinného dolu. Odpůrci těžby vyčíslili její vedlejší finanční náklady na 269 miliard. Do této sumy započítali výdaje na léčení nemocí způsobených škodlivinami ze spalování uhlí při výrobě elektrické energie a tepla, jako jsou srdeční choroby a chronická bronchitida, výplatu nemocenských dávek nebo škody na zemědělské půdě a na krajině. Ústecký hejtman Oldřich Bubeníček (KSČM) naopak věří, že by pracovní místa přibyla. Kraj má přes 60 tisíc nezaměstnaných. V případě završení těžby se hejtman děsí "sociální katastrofy". Podobně rázně hovoří odbory. Řeč s nimi na včerejší tripartitě našli i podnikatelé - v uhlí vidí naše bohatství, jež nemá ležet ladem. Zástupci státu mají ale pochyby. ANO, lidovci a podstatné části ČSSD se prolomení limitů nezamlouvá. Ministr financí Andrej Babiš, který mívá přesně zmapováno, co je efektivní a populární, o něm nechce slyšet a upozorňuje, že další hnědé uhlí nepotřebujeme - energii nakonec stát vyváží. Premiér Bohuslav Sobotka odmítá bourat obydlí a usiluje o kompromisní řešení, jež uspokojí starousedlíky, Lidovému domu blízké odboráře i teplárenský průmysl. Od ministra průmyslu Jana Mládka natvrdo zaznělo, že pro teplárenství je uhlí nepostradatelné. Straší, že zachování limitů by nás mohlo stát vyšší ceny tepla. Prostor pro upachtěnou dohodu však existuje. Na stole jsou čtyři alternativy.
Statements
ID
Text
all
current
filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna


Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title
#
Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT
0
Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT
0
Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT
0
Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT
0
Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT
0
Feb 4, 2015
Na těžební limity je jiný pohled z Horního Jiřetína či Litvínova, jiný z Ústí nad Labem a jiný z Prahy. Proto jsme svědky debaty, která má vertikální i horizontální linii, a vstupují do ní akcenty politické, ekonomické, sociální i ekologické. Místní už kdysi řekli v lokálních referendech "ne" a od té doby na radnicích sílí subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního Jiřetína Vladimír Buřt (SZ) se snaží oslabit hlavní trumf protistrany, jež tvrdí, že rozšíření těžby pomůže zaměstnanosti. Podle něj by v katastru obce bylo ohroženo asi osm set pracovních míst a na povro ily i tři stovky zaměstnanců místního hlubinného dolu. li její vedlejší finanční náklady na 269 miliard. Do této sumy započítali výdaje na léčení nemocí způsobených škodlivinami ze spalování uhlí při výrobě elektrické energie a tepla, jako jsou srdeční choroby a chronická bronchitida, výplatu nemocenských dávek nebo škody na zemědělské půdě a na krajině. Ústecký hejtman Oldřich Bubeníček (KSČM) naopak věří, že by pracovní místa přibyla. Kraj má přes 60 tisíc nezaměstnaných. V případě završení těžby se hejtman děsí "sociální katastrofy". Podobně rázně hovoří odbory. Řeč s nimi na včerejší tripartitě našli i podnikatelé - v uhlí vidí naše bohatství, jež nemá ležet ladem. Zástupci státu mají ale pochyby. ANO, lidovci a podstatné části ČSSD se prolomení limitů nezamlouvá. Ministr financí Andrej Babiš, který mívá přesně zmapováno, co je efektivní a populární, o něm nechce slyšet a upozorňuje, že další hnědé uhlí nepotřebujeme - energii nakonec stát vyváží. Premiér Bohuslav Sobotka odmítá bourat obydlí a usiluje o kompromisní řešení, jež uspokojí starousedlíky, Lidovému domu blízké odboráře i teplárenský průmysl. Od ministra průmyslu Jana Mládka natvrdo zaznělo, že pro teplárenství je uhlí nepostradatelné. Straší, že zachování limitů by nás mohlo stát vyšší ceny tepla. Prostor pro upachtěnou dohodu však existuje. Na stole jsou čtyři alternativy.
Format as DNA Statement
Format as Annotation
Statements
ID
Text
all
current
filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna


Discourse Network Analyzer
File Document Export Settings
220
14
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title # Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 1 Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 0 Feb 4, 2015
Na těžební limity je jiný pohled z Horního Jiřetína či Litvínova, jiný z Ústí nad Labem a jiný z Prahy. Proto jsme svědky debaty, která má vertikální i horizontální linii, a vstupují do ní akcenty politické, ekonomické, sociální i ekologické. Místní už kdysi řekli v lokálních referendech "ne" a od té doby na radnicích sílí subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního Jiřetína Vladimír Buřt (SZ) se snaží oslabit hlavní trumf protistrany, jež tvrdí, že rozšíření těžby pomůže zaměst lo ohroženo asi osm set pracovních míst a na vky zaměstnanců místního hlubinného jší finanční náklady na 269 miliard. í nemocí způsobených škodlivinami ze spal a tepla, jako jsou srdeční choroby a chro dávek nebo škody na zemědělské půdě a na krajině. Ústecký hejtman Oldřich Bubeníček (KSČM) naopak věří, že by pracovní místa přibyla. Kraj má přes 60 tisíc nezaměstnaných. V případě završení těžby se hejtman děsí "sociální katastrofy". Podobně rázně hovoří odbory. Řeč s nimi na včerejší tripartitě našli i podnikatelé - v uhlí vidí naše bohatství, jež nemá ležet ladem. Zástupci státu mají ale pochyby. ANO, lidovci a podstatné části ČSSD se prolomení limitů nezamlouvá. Ministr financí Andrej Babiš, který mívá přesně zmapováno, co je efektivní a populární, o něm nechce slyšet a upozorňuje, že další hnědé uhlí nepotřebujeme - energii nakonec stát vyváží. Premiér Bohuslav Sobotka odmítá bourat obydlí a usiluje o kompromisní řešení, jež uspokojí starousedlíky, Lidovému domu blízké odboráře i teplárenský průmysl. Od ministra průmyslu Jana Mládka natvrdo zaznělo, že pro teplárenství je uhlí nepostradatelné. Straší, že zachování limitů by nás mohlo stát vyšší ceny tepla. Prostor pro upachtěnou dohodu však existuje. Na stole jsou čtyři alternativy.
DNA Statement ID: 1 start: 439 end: 529
person Vladimir Burt
organization Strana Zelenych
concept zamestnanost
agreement
Statements
ID Text
1 snaží oslabit hlavní trumf pro...
all current filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna

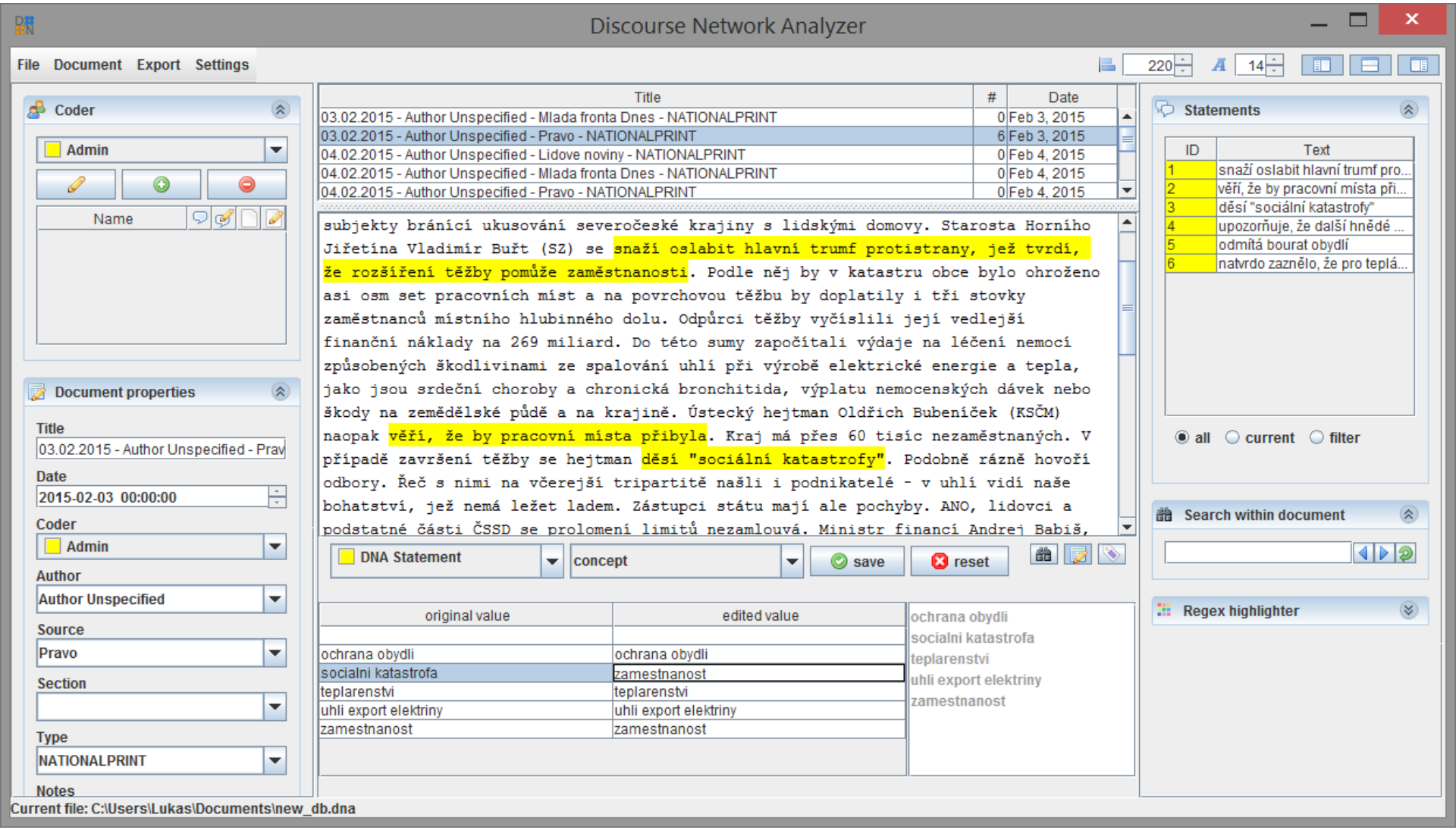

Discourse Network Analyzer
File Document Export Settings
220
14
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title # Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 2 Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 0 Feb 4, 2015
Labem a jiný z Prahy. Proto jsme svědky debaty, která má vertikální i horizontální
linii, a vstupují do ní akcenty politické, ekonomické, sociální i ekologické.
Místní už kdysi řekli v lokálních referendech "ne" a od té doby na radnicích sílí
subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního
Jiřetína Vladimír Bušt (SZ) se snaží oslabit hlavní trumf protistrany, jež tvrdí,
že rozšíření těžby pomůže zaměstnanosti. Podle něj by v katastru obce bylo ohroženo
asi osm set pracovních míst a na povrchovou těžbu by doplatily i tři stovky
zaměstnanců místního hlubinného dolu. Odpůrci těžby vyčíslili její vedlejší
finanční náklady na 269 miliard. Do této sumy započítali výdaje na léčení nemocí
způsobených škodlivinami ze spalování uhlí při výrobě elektrické energie a tepla,
jako jsou srdeční choroby a chronická bronchitida, výplatu nemocenských dávek nebo
škody na zemědělské půdě a na krajině. Ústecký hejtman Oldřich Bubeníček (KSČM)
naopak
případě
odbory.
bohatstv
podstatn
který mí
upozorňu
katastrofy". Podobně rázně hovoří
podnikatelé - v uhlí vidí naše
ají ale pochyby. ANO, lidovci a
rá. Ministr financí Andrej Babiš,
pulární, o něm nechce slyšet a
ergii nakonec stát vyváží.
Premiér Bohuslav Sobotka odmítá bourat obydlí a usiluje o kompromisní řešení, jež
uspokojí starousedlíky, Lidovému domu blízké odboráře i teplárenský průmysl. Od
ministra průmyslu Jana Mládka natvrdo zaznělo, že pro teplárenství je uhlí
nepostradatelné. Straší, že zachování limitů by nás mohlo stát vyšší ceny tepla.
Prostor pro upachtěnou dohodu však existuje. Na stole jsou čtyři alternativy,
Mládek pokládá za reálné dvě: limity budou upraveny jen na dole Bílina (což
DNA Statement ID: 2 start: 1059 end: 1093
person Oldrich Bubenicek
organization Hejtman UnL
concept
agreement zamestnanost
Statements
ID Text
1 snaží oslabit hlavní trumf pro...
2 věří, že by pracovní místa při...
all current filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna

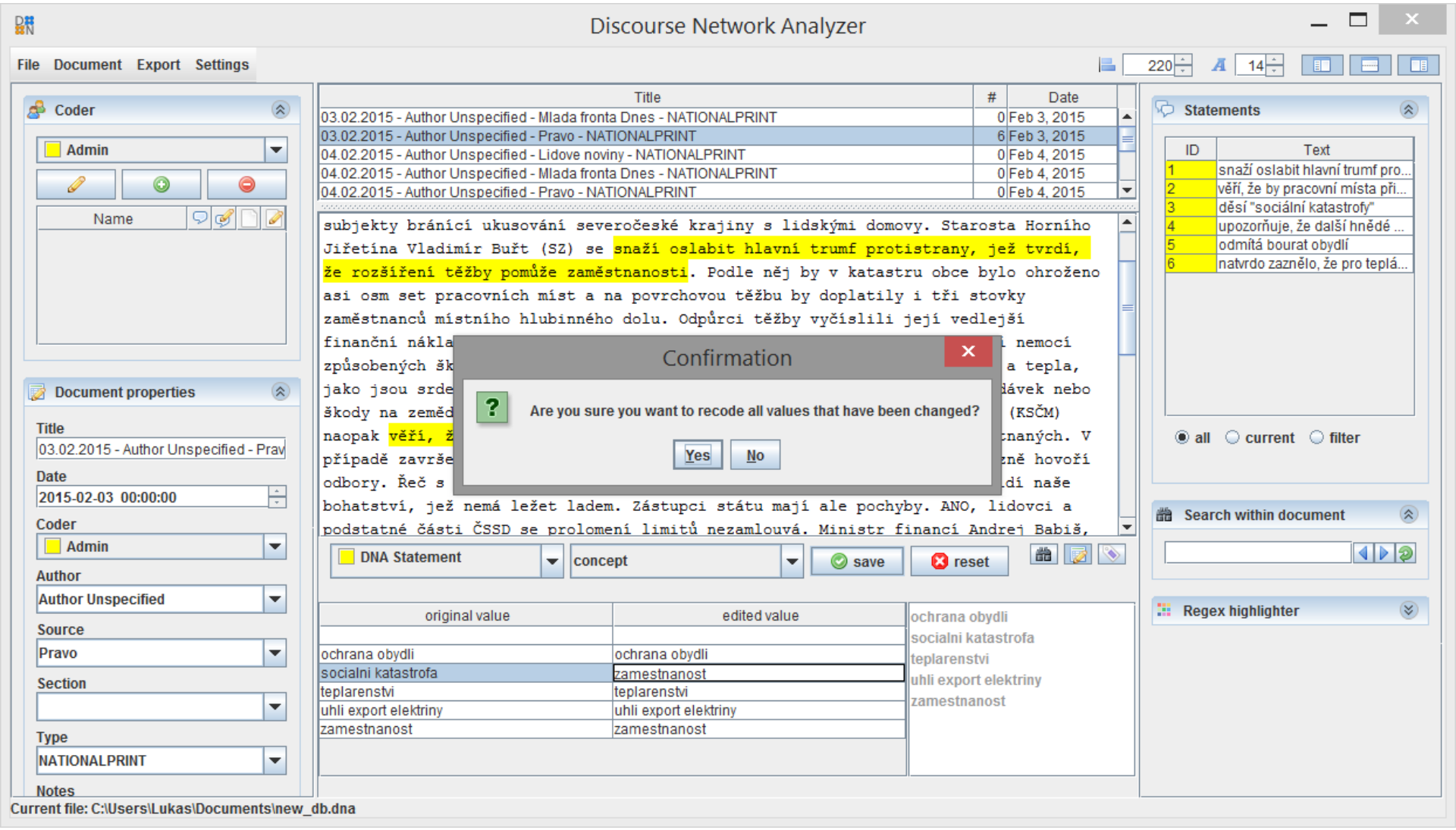

Discourse Network Analyzer
File Document Export Settings
220
14
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title # Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 3 Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 0 Feb 4, 2015
linii, a vstupují do ní akcenty politické, ekonomické, sociální i ekologické.
Místní už kdysi řekli v lokálních referendech "ne" a od té doby na radnicích sílí
subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního
Jiřetína Vladimír Bufť (SZ) se snaží oslabit hlavní trumf protistrany, jež tvrdí,
že rozšíření těžby pomůže zaměstnanosti. Podle něj by v katastru obce bylo ohroženo
asi osm set pracovních míst a na povrchovou těžbu by doplatily i tři stovky
zaměstnanců místního hlubinného dolu. Odpůrci těžby vyčíslili její vedlejší
finanční náklady na 269 miliard. Do této sumy započítali výdaje na léčení nemocí
způsobených škodlivinami ze spalování uhlí při výrobě elektrické energie a tepla,
jako jsou srdeční choroby a chronická bronchitida, výplatu nemocenských dávek nebo
škody na zemědělské půdě a na krajině. Ústecký hejtman Oldřich Bubeníček (KSČM)
naopak věří, že by pracovní místa přibyla. Kraj má přes 60 tisíc nezaměstnaných. V
případě završení těžby se hejtman děsí "sociální katastrofy". Podobně rázně hovoří
odbory. Řeč s nimi na včerejší trip
bohatství, jež nemá ležet ladem. Zá
podstatné části ČSSD se prolomení l
který mívá přesně zmapováno, co je
upozorňuje, že další hnědé uhlí nep
Premiér Bohuslav Sobotka odmítá bou
uspokojí starousedlíky, Lidovému domu blízké odboráře i teplárenský průmysl. Od
ministra průmyslu Jana Mládka natvrdo zaznělo, že pro teplárenství je uhlí
nepostradatelné. Straší, že zachování limitů by nás mohlo stát vyšší ceny tepla.
Prostor pro upachtěnou dohodu však existuje. Na stole jsou čtyři alternativy,
Mládek pokládá za reálné dvě: limity budou upraveny jen na dole Bílina (což
připouští i ministr životního prostředí Richard Brabec z ANO), nebo na Bílině a
DNA Statement ID: 3 start: 1169 end: 1195
person
organization
Oldrich Bubenicek
Vladimir Burt
concept
agreement
Statements
ID Text
1 snaží oslabit hlavní trumf pro...
2 věří, že by pracovní místa při...
3 děsí "sociální katastrofy"
all current filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna


Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title
Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 6 Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 0 Feb 4, 2015
subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního Jiřetína Vladimír Buřt (SZ) se snaží oslabit hlavní trumf protistrany, jež tvrdí, že rozšíření těžby pomůže zaměstnanosti. Podle něj by v katastru obce bylo ohroženo asi osm set pracovních míst a na povrchovou těžbu by doplatily i tři stovky zaměstnanců místního hlubinného dolu. Odpůrci těžby vyčíslili její vedlejší finanční náklady na 269 miliard. Do této sumy započítali výdaje na léčení nemocí způsobených škodlivinami ze spalování uhlí při výrobě elektrické energie a tepla, jako jsou srdeční choroby a chronická bronchitida, výplatu nemocenských dávek nebo škody na zemědělské půdě a na krajině. Ústecký hejtman Oldřich Bubeníček (KSČM) naopak věří, že by pracovní místa přibyla. Kraj má přes 60 tisíc nezaměstnaných. V případě završení těžby se hejtman děsí "sociální katastrofy". Podobně rázně hovoří odbory. Řeč s nimi na včerejší tripartitě našli i podnikatelé - v uhlí vidí naše bohatství, jež nemá ležet ladem. Zástupci státu mají ale pochyby. ANO, lidovci a podstatné části ČSSD se prolomení limitů nezamlouvá. Ministr financí Andrej Babiš, který mívá přesně zmapováno, co je efektivní a populární, o něm nechce slyšet a upozorňuje, že další hnědé uhlí nepotřebujeme - energii nakonec stát vyváží. Premiér Bohuslav Sobotka odmítá bourat obydlí a usiluje o kompromisní řešení, jež uspokojí starousedlíky, Lidovému domu blízké odboráře i teplárenský průmysl. Od ministra průmyslu Jana Mládka natvrdo zaznělo, že pro teplárenství je uhlí nepostradatelné. Straší, že zachování limitů by nás mohlo stát vyšší ceny tepla. Prostor pro upachtěnou dohodu však existuje. Na stole jsou čtyři alternativy, Mládek pokládá za reálné dvě: limity budou upraveny jen na dole Bílina (což připouští i ministr životního prostředí Richard Brabec z ANO), nebo na Bílině a částečně na dole ČSA - padl by jih Horního Jiřetína čítající 170 domů (to doporučuje Mládek). Okrajovými jsou varianty úplného prolomení limitů (zbourání
Statements
ID Text
1 snaží oslabit hlavní trumf pro...
2 věří, že by pracovní místa při...
3 děsí "sociální katastrofy"
4 upozorňuje, že další hnědé ...
5 odmítá bourat obydlí
6 natvrdo zaznělo, že pro teplá...
all
current
filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna

Recoding

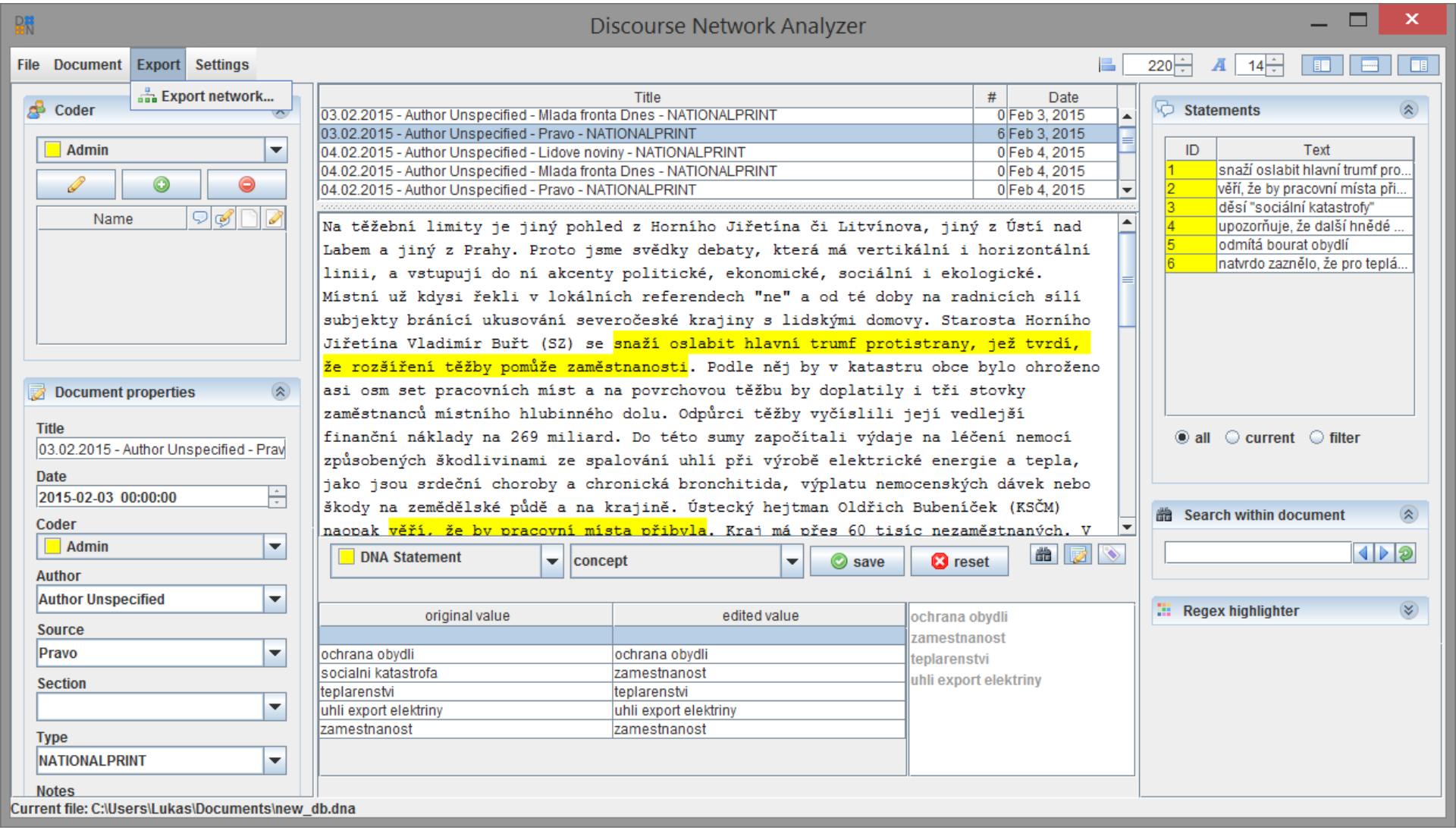
- Open coding as first step
- Establishment of relations between codes (axial coding)
  - Adjustment of original codes' labels
  - Reduction of dimensions
- Any variable can be recoded
  - From original variable value to target value

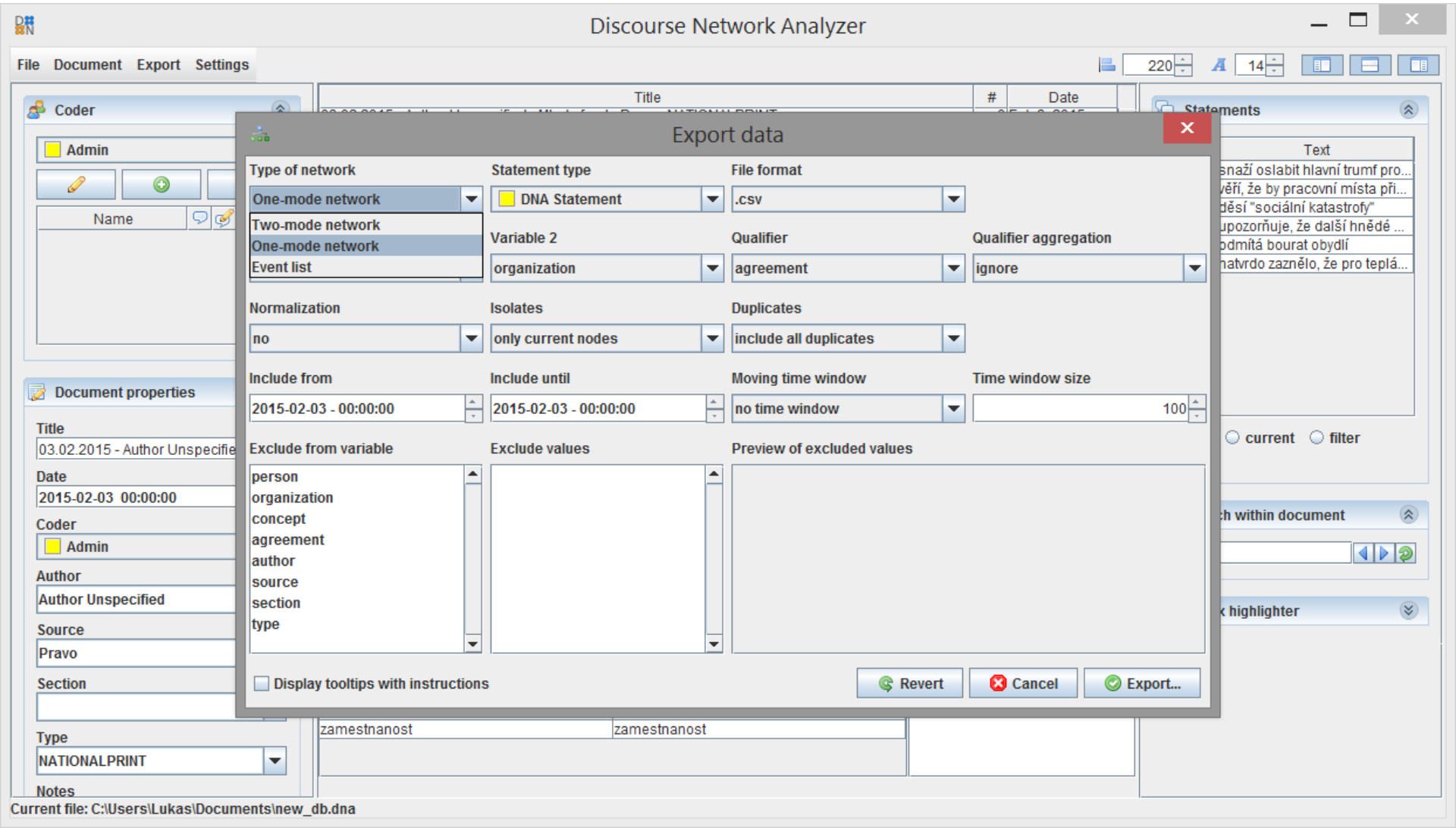

Discourse Network Analyzer
File
Document
Export
Settings
220
14
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title
#
Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT
0
Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT
6
Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT
0
Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT
0
Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT
0
Feb 4, 2015
subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního Jiřetína Vladimír Buřt (SZ) se snaží oslabit hlavní trumf protistrany, jež tvrdí, že rozšíření těžby pomůže zaměstnanosti. Podle něj by v katastru obce bylo ohroženo asi osm set pracovních míst a na povrchovou těžbu by doplatily i tři stovky zaměstnanců místního hlubinného dolu. Odpůrci těžby vyčíslili její vedlejší finanční náklady na 269 miliard. Do této sumy započítali výdaje na léčení nemocí způsobených škodlivinami ze spalování uhlí při výrobě elektrické energie a tepla, jako jsou srdeční choroby a chronická bronchitida, výplatu nemocenských dávek nebo škody na zemědělské půdě a na krajině. Ústecký hejtman Oldřich Bubeníček (KSČM) naopak věří, že by pracovní místa přibyla. Kraj má přes 60 tisíc nezaměstnaných. V případě završení těžby se hejtman děsí "sociální katastrofy". Podobně rázně hovoří odbory. Řeč s nimi na včerejší tripartitě našli i podnikatelé - v uhlí vidí naše bohatství, jež nemá ležet ladem. Zástupci státu mají ale pochyby. ANO, lidovci a podstatné části ČSSD se prolomení limitů nezamlouvá. Ministr financí Andrej Babiš,
DNA Statement
concept
save
reset
original value
edited value
ochrana obydli
ochrana obydli
socialni katastrofa
zamestnanost
teplarenstvi
teplarenstvi
uhli export elektriny
uhli export elektriny
zamestnanost
ochrana obydli
socialni katastrofa
teplarenstvi
uhli export elektriny
zamestnanost
Statements
ID
Text
1
snaží oslabit hlavní trumf pro...
2
věří, že by pracovní místa při...
3
děsí "sociální katastrofy"
4
upozorňuje, že další hnědé ...
5
odmítá bourat obydlí
6
natvrdo zaznělo, že pro teplá...
all
current
filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna

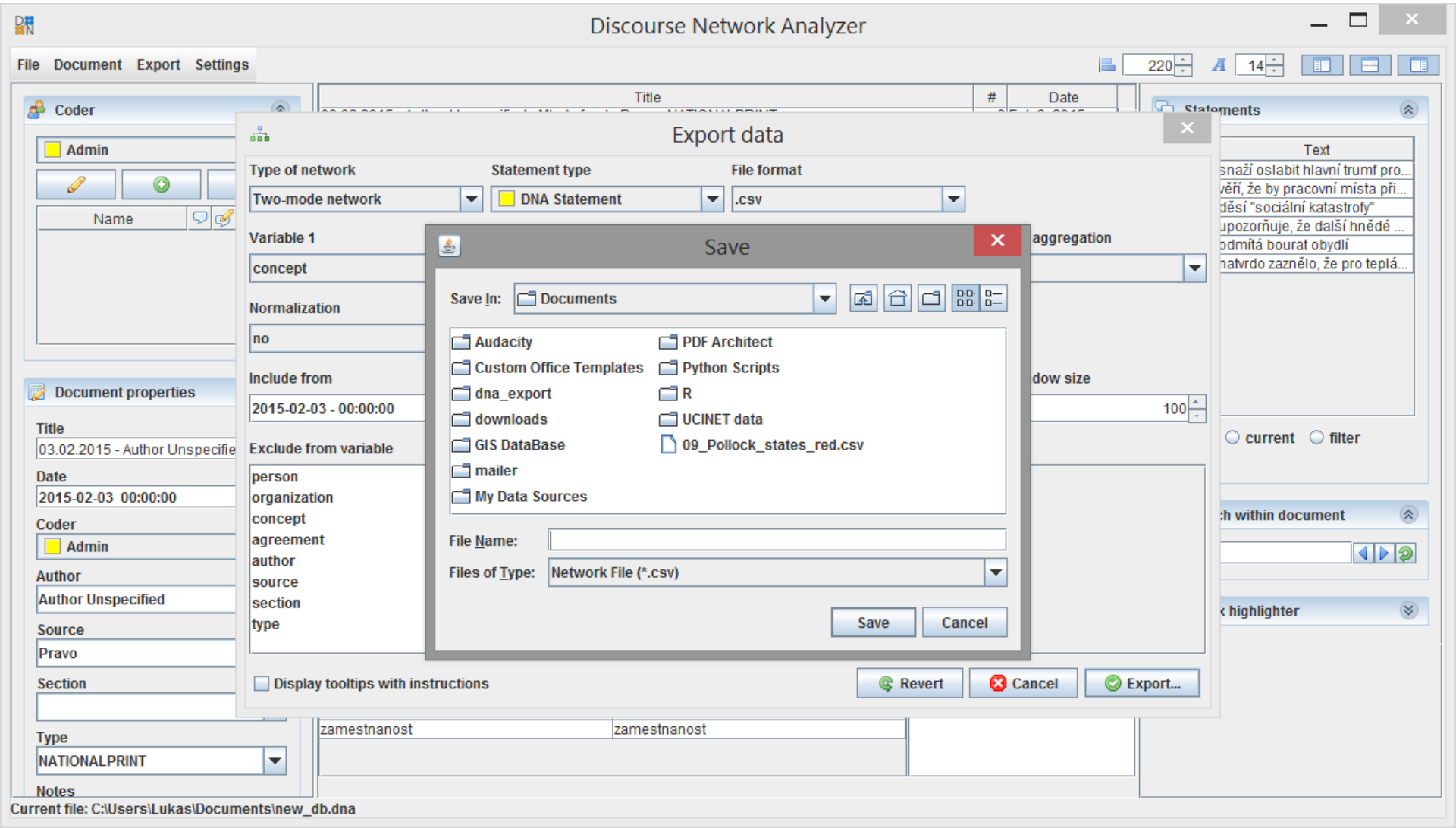

Discourse Network Analyzer
File Document Export Settings
220
14
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title # Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 6 Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 0 Feb 4, 2015
subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního Jiřetína Vladimír Buřt (SZ) se snaží oslabit hlavní trumf protistrany, jež tvrdí, že rozšíření těžby pomůže zaměstnanosti. Podle něj by v katastru obce bylo ohroženo asi osm set pracovních míst a na povrchovou těžbu by doplatily i tři stovky zaměstnanců místního hlubinného dolu. Odpůrci těžby vyčíslili její vedlejší
Confirmation
Are you sure you want to recode all values that have been changed?
Yes
No
DNA Statement
concept
save
reset
original value edited value
ochrana obydli ochrana obydli
socialni katastrofa zamestnanost
teplarenstvi teplarenstvi
uhli export elektriny uhli export elektriny
zamestnanost zamestnanost
ochrana obydli
socialni katastrofa
teplarenstvi
uhli export elektriny
zamestnanost
Statements
ID Text
1 snaží oslabit hlavní trumf pro...
2 věří, že by pracovní místa při...
3 děsí "sociální katastrofy"
4 upozorňuje, že další hnědé ...
5 odmítá bourat obydlí
6 natvrdo zaznělo, že pro teplá...
all current filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna

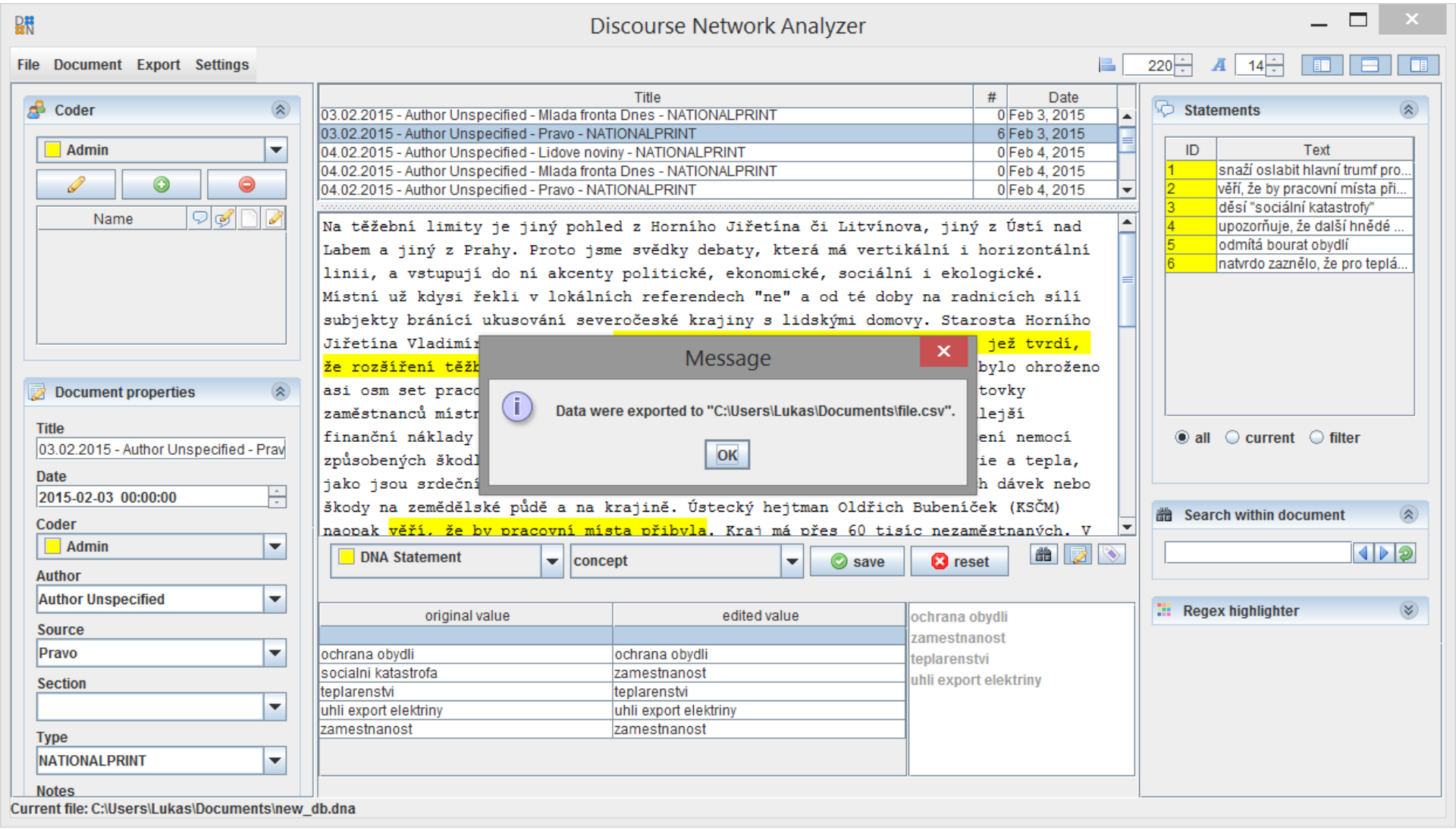
Data export

- Versatile data export possibility
  - One-mode network – values of single variable
  - Two-mode network – matrix of values
- Data export formats
  - CSV
  - DL (network analysis software UCInet)
  - GRAPHML (network visualization software Visone)


Discourse Network Analyzer
File
Document
Export
Settings
Export network...
220
14
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title
#
Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 6 Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 0 Feb 4, 2015
Na těžební limity je jiný pohled z Horního Jiřetína či Litvínova, jiný z Ústí nad Labem a jiný z Prahy. Proto jsme svědky debaty, která má vertikální i horizontální linii, a vstupují do ní akcenty politické, ekonomické, sociální i ekologické. Místní už kdysi řekli v lokálních referendech "ne" a od té doby na radnicích sílí subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního Jiřetína Vladimír Bušt (SZ) se snaží oslabit hlavní trumf protistrany, jež tvrdí, že rozšíření těžby pomůže zaměstnanosti. Podle něj by v katastru obce bylo ohroženo asi osm set pracovních míst a na povrchovou těžbu by doplatily i tři stovky zaměstnanců místního hlubinného dolu. Odpůrci těžby vyčíslili její vedlejší finanční náklady na 269 miliard. Do této sumy započítali výdaje na léčení nemocí způsobených škodlivinami ze spalování uhlí při výrobě elektrické energie a tepla, jako jsou srdeční choroby a chronická bronchitida, výplatu nemocenských dávek nebo škody na zemědělské půdě a na krajině. Ústecký hejtman Oldřich Bubeníček (KSČM) naopak věří, že by pracovní místa přibyla. Kraj má přes 60 tisíc nezaměstnaných. V
DNA Statement
concept
save
reset
original value
edited value
ochrana obydli ochrana obydli
socialni katastrofa zamestnanost
teplarenstvi teplarenstvi
uhli export elektriny uhli export elektriny
zamestnanost zamestnanost
ochrana obydli
zamestnanost
teplarenstvi
uhli export elektriny
Statements
ID
Text
1 snaží oslabit hlavní trumf pro...
2 věří, že by pracovní místa při...
3 děsí "sociální katastrofy"
4 upozorňuje, že další hnědé ...
5 odmítá bourat obydlí
6 natvrdo zaznělo, že pro teplá...
all
current
filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna


Discourse Network Analyzer
File
Document
Export
Settings
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecifie
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title
#
Date
zamestnanost
zamestnanost
Statements
Text
snaží oslabit hlavní trumf pro...
věří, že by pracovní místa při...
děsí "sociální katastrofy"
upozorňuje, že další hnědé ...
odmítá bourat obydlí
natvrdo zaznělo, že pro teplá...
current
filter
h within document
x highlighter
Export data
Type of network
One-mode network
Two-mode network
One-mode network
Event list
Statement type
DNA Statement
File format
.csv
Variable 2
organization
Qualifier
agreement
Qualifier aggregation
ignore
Normalization
no
Isolates
only current nodes
Duplicates
include all duplicates
Include from
2015-02-03 - 00:00:00
Include until
2015-02-03 - 00:00:00
Moving time window
no time window
Time window size
100
Exclude from variable
person
organization
concept
agreement
author
source
section
type
Exclude values
Preview of excluded values
Display tooltips with instructions
Revert
Cancel
Export...
Current file: C:\Users\Lukas\Documents\new_db.dna

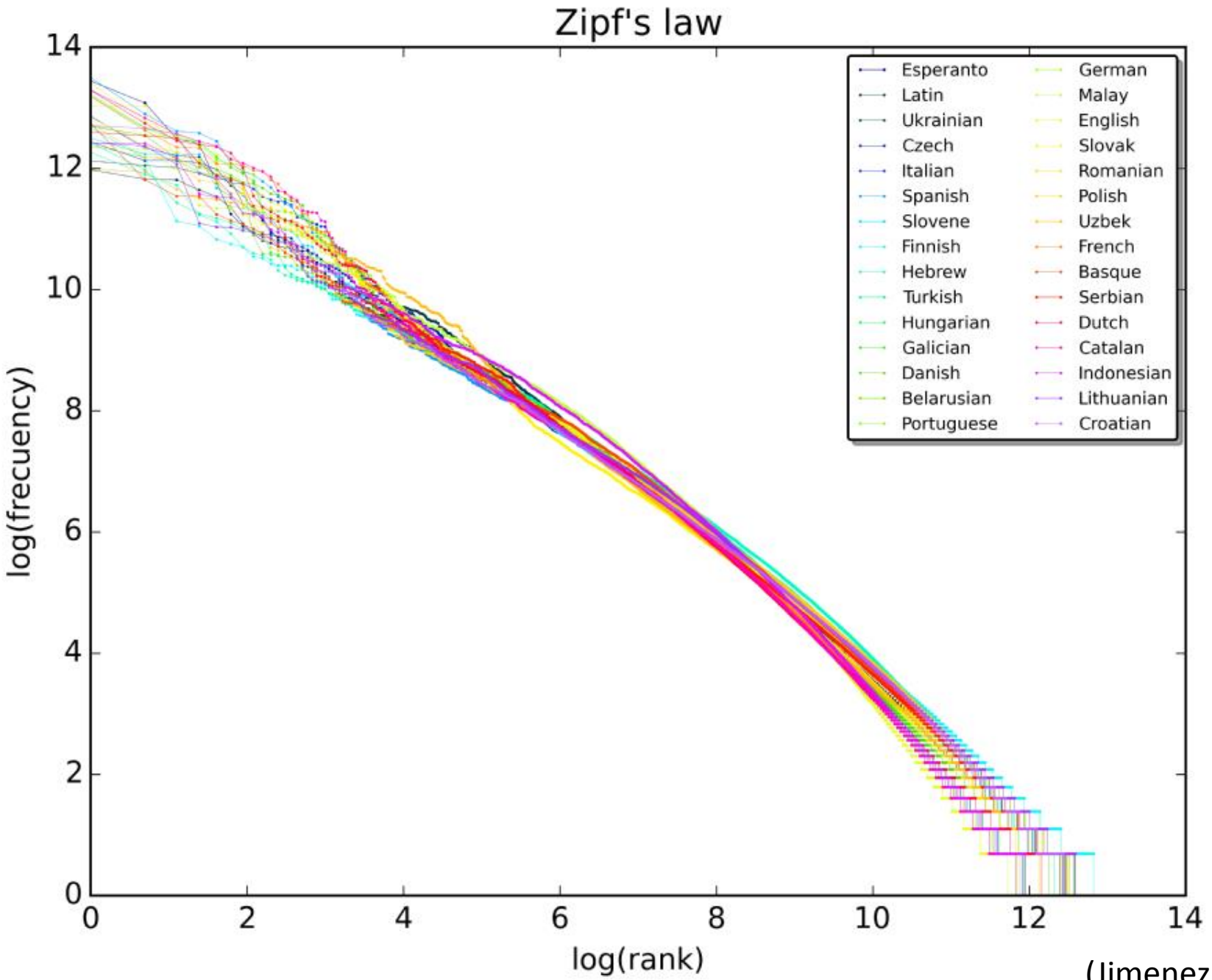

Discourse Network Analyzer
File Document Export Settings
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecifie
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title # Date
zamestnanost zamestnanost
Statements
Text
Export data
Type of network
Two-mode network
Statement type
DNA Statement
File format
.csv
Variable 1
concept
Normalization
no
Include from
2015-02-03 - 00:00:00
Exclude from variable
person
organization
concept
agreement
author
source
section
type
aggregation
dow size
100
current
filter
Save
Save In:
Documents
Audacity
Custom Office Templates
dna_export
downloads
GIS DataBase
mailer
My Data Sources
PDF Architect
Python Scripts
R
UCINET data
09_Pollock_states_red.csv
File Name:
Files of Type:
Network File (*.csv)
Save
Cancel
Display tooltips with instructions
Revert
Cancel
Export...
Current file: C:\Users\Lukas\Documents\new_db.dna


Discourse Network Analyzer
File Document Export Settings
Coder
Admin
Name
Document properties
Title
03.02.2015 - Author Unspecified - Prav
Date
2015-02-03 00:00:00
Coder
Admin
Author
Author Unspecified
Source
Pravo
Section
Type
NATIONALPRINT
Notes
Title # Date
03.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 3, 2015
03.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 6 Feb 3, 2015
04.02.2015 - Author Unspecified - Lidove noviny - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Mlada fronta Dnes - NATIONALPRINT 0 Feb 4, 2015
04.02.2015 - Author Unspecified - Pravo - NATIONALPRINT 0 Feb 4, 2015
Na těžební limity je jiný pohled z Horního Jiřetína či Litvínova, jiný z Ústí nad Labem a jiný z Prahy. Proto jsme svědky debaty, která má vertikální i horizontální linii, a vstupují do ní akcenty politické, ekonomické, sociální i ekologické. Místní už kdysi řekli v lokálních referendech "ne" a od té doby na radnicích sílí subjekty bránící ukusování severočeské krajiny s lidskými domovy. Starosta Horního Jiřetína Vladimí
jež tvrdí,
že rozšíření těžb
bylo ohroženo
asi osm set pracc
tovky
zaměstnanců místr
lejší
finanční náklady
ení nemocí
způsobených škodl
ie a tepla,
jako jsou srdeční
h dávek nebo
škody na zemědělské půdě a na krajině. Ústecký hejtman Oldřich Bubeníček (KSČM) naopak věří, že by pracovní místa přibyla. Kraj má přes 60 tisíc nezaměstnaných. V
Message
Data were exported to "C:\Users\Lukas\Documents\file.csv".
OK
DNA Statement
concept
save
reset
original value
edited value
ochrana obydli ochrana obydli
socialni katastrofa zamestnanost
teplarenstvi teplarenstvi
uhli export elektriny uhli export elektriny
zamestnanost zamestnanost
ochrana obydli
zamestnanost
teplarenstvi
uhli export elektriny
Statements
ID Text
1 snaží oslabit hlavní trumf pro...
2 věří, že by pracovní místa při...
3 děsí "sociální katastrofy"
4 upozorňuje, že další hnědé ...
5 odmítá bourat obydlí
6 natvrdo zaznělo, že pro teplá...
all current filter
Search within document
Regex highlighter
Current file: C:\Users\Lukas\Documents\new_db.dna

What next?

Quantitative TA

Zipf law



(Jimenez, 2015)

Corpus

Corpus

- Decision over document unitizing
- Decision over sampling
  - Does 5M texts provide more information than 15k?
  - Random vs. non-random sampling
- Inclusion of metadata – additional information
  - Author
  - Time and date
  - Source (e.g. media/newspaper)
  - …

Words and content

Words and content

- Some words used to convey meaning, other are used functionally to allow meaningful language
  - Nouns, verbs, adjectives, pronouns
  - Stopwords – make sense only when connected with other terms
- Depends on task at hand
  - In some cases, it is justified to drop them
  - In others, these words are important

Text pre-processing

- Considerations over pre-processing
  - Dropping sparse terms
  - Dropping most frequent terms
  - Dropping “stopwords”
  - Dropping numerals, punctuation, …
  - Dropping time and place information
  - …
- Method dependent
- Sometimes affects results (topic modeling)

Text pre-processing

- Stemming/lemmatization
  - Disposal of grammatical features of text
    - Dictionary-based
    - Rules-based
  - Both introduce some error into the corpus
- Lemmatization
  - Identification of lemmas (lexemes) of the words - transformation to lemmas
- Stemming
  - Stripping the word of prefixes or suffixes, leaving only word stems

Lemmatization and stemming

"This was the most tranquil presidential address. President's approach was very relaxed."

- Lemmatization

"This be the most tranquil presidential address. President approach be very relax."

- Stemming

"This be the most tranquil presidenti address.

Presid approach be veri relax."

Bag of words

Bag of words

- The quick brown fox jumps over the lazy dog

Word	Occurrence
brown	1
dog	1
fox	1
jumps	1
lazy	1
over	1
quick	1
the	2

Document-feature matrix

- Matrix – most methods based on this
  - 1st dim – Features/Tokens (words, phrases, …)
  - 2nd dim – Documents/units
  - Cells – frequency of tokens in documents
    - Boolean – Present vs. Not present (1/0)
    - Weighted
      - Absolute frequency (how many times word occur in document)
      - TF-IDF
- Grows large easily
  - 500 documents * 1k unique tokens = 0.5M cells
- Usually very sparse
  - Most of cells are empty – contain 0

Document-feature matrix

	2003-2004-cz	2004-2005-pl	2005-2006-hu	2006-2007-sk	2007-2008-cz	Sum
agriculture	3	6	2	5	3	**19**
aim	4	2	7	12	6	**31**
area	11	8	8	28	26	**81**
base	1	2	2	2	5	**12**
border	5	9	9	3	3	**29**
central	2	3	6	3	5	**19**
cohesion	3	1	7	4	4	**19**
commission	2	7	3	2	4	**18**
common	10	9	17	8	17	**61**
community	2	2	3	3	6	**16**
concern	9	13	12	18	6	**58**

Co-occurrence

Co-occurrence

- The quick **brown fox** jumps over the lazy dog.
- **Brown dog** sleeps well.

Feature	Document 1	Document 2
brown	1	1
dog	1	1
fox	1	
jumps	1	
lazy	1	
over	1	
quick	1	
sleeps		1
the	2	
well		1

Co-locations and N-grams

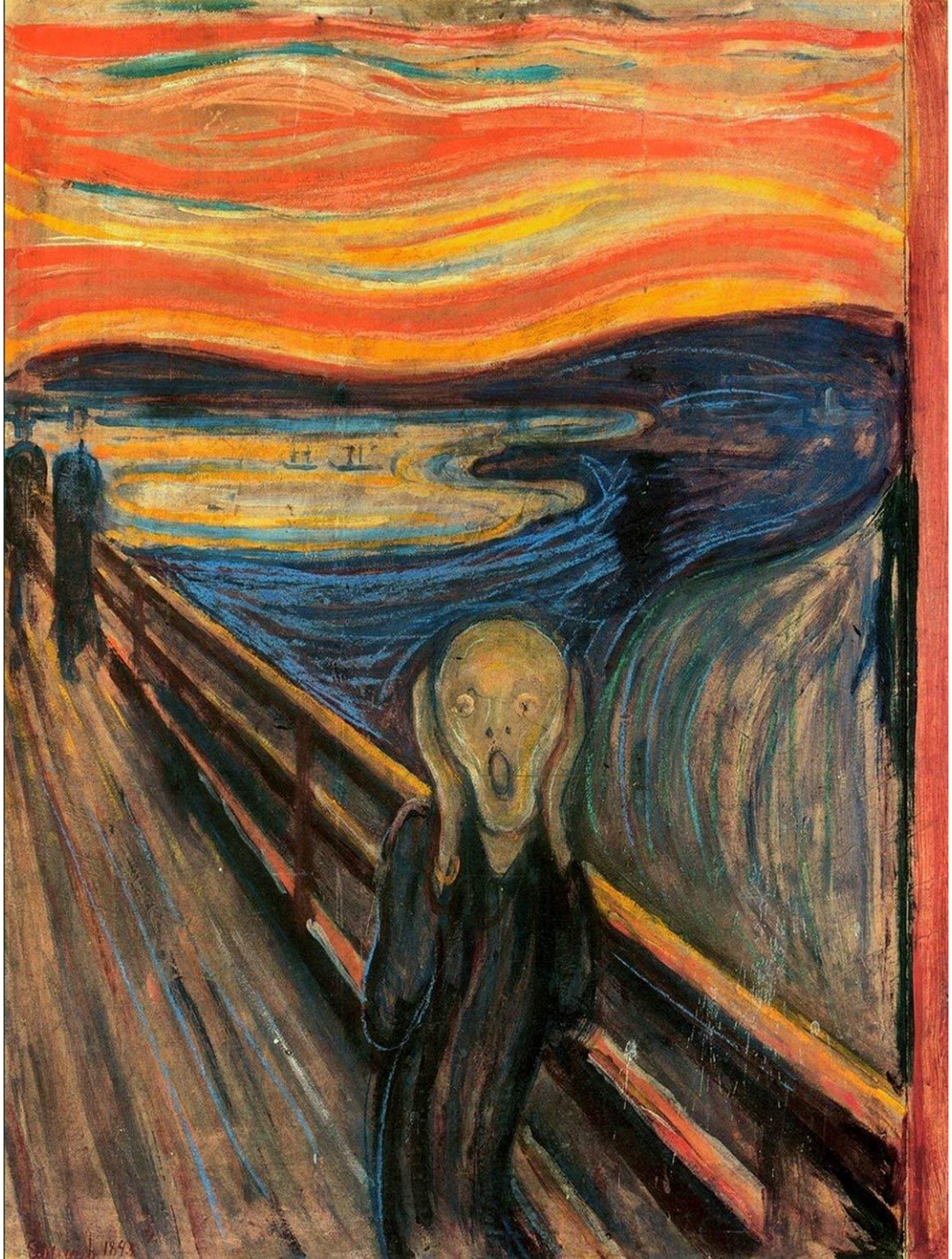
Co-locations/n-grams

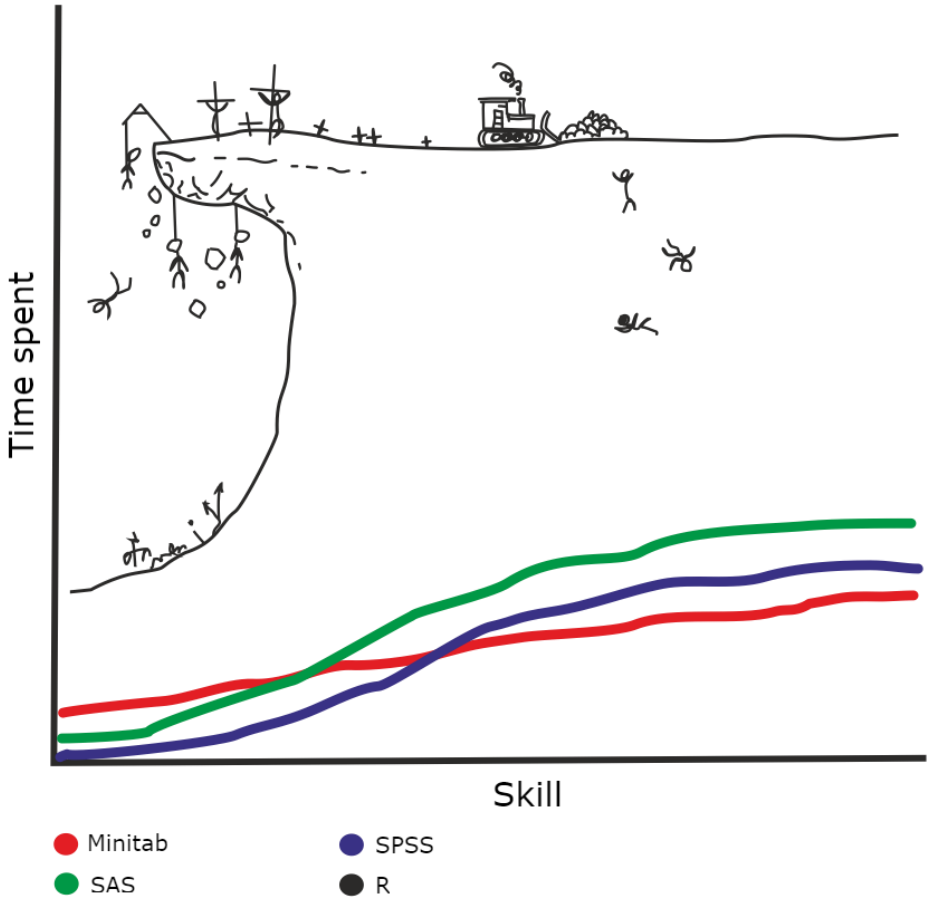
- Co-locations
  - Established phrases, usually occur together
  - Provide little information over text
    - Ministry of the Environment
    - European Union
    - prime minister
- N-grams
  - Phrases which are not established, but occur together in text
  - Provide insights
    - Crooked Hillary

„Be careful what is a result and what is just a residue of your data choices“

Jana Diesner, 2018

R basics




Time spent
Skill
Minitab
SAS
SPSS
R

R is a language

- Any programming language is just very **condensed and formalized** speech
- Understand and formulate the **process is key**
- Scripting is just a matter of knowing right expressions

Many resources out there

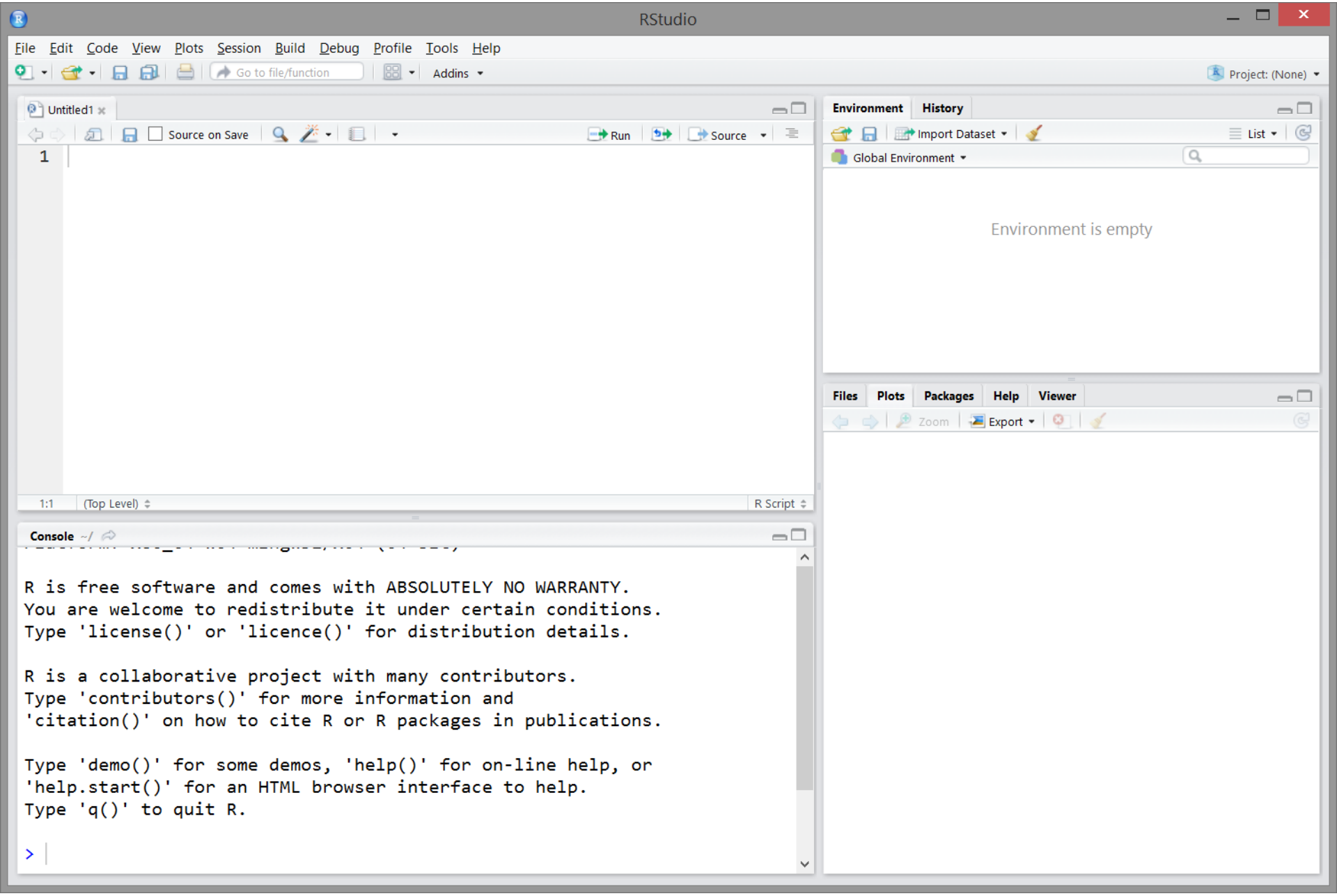
- R package / library manuals
- R site: http://cran.r-project.org
- community forums:
  - http://stackoverflow.com
  - http://www.statmethods.net
  - http://www.r-bloggers.com
- Youtube videos: https://www.youtube.com/watch?v=qHfSTRNg6jE
- googling (often fastest)

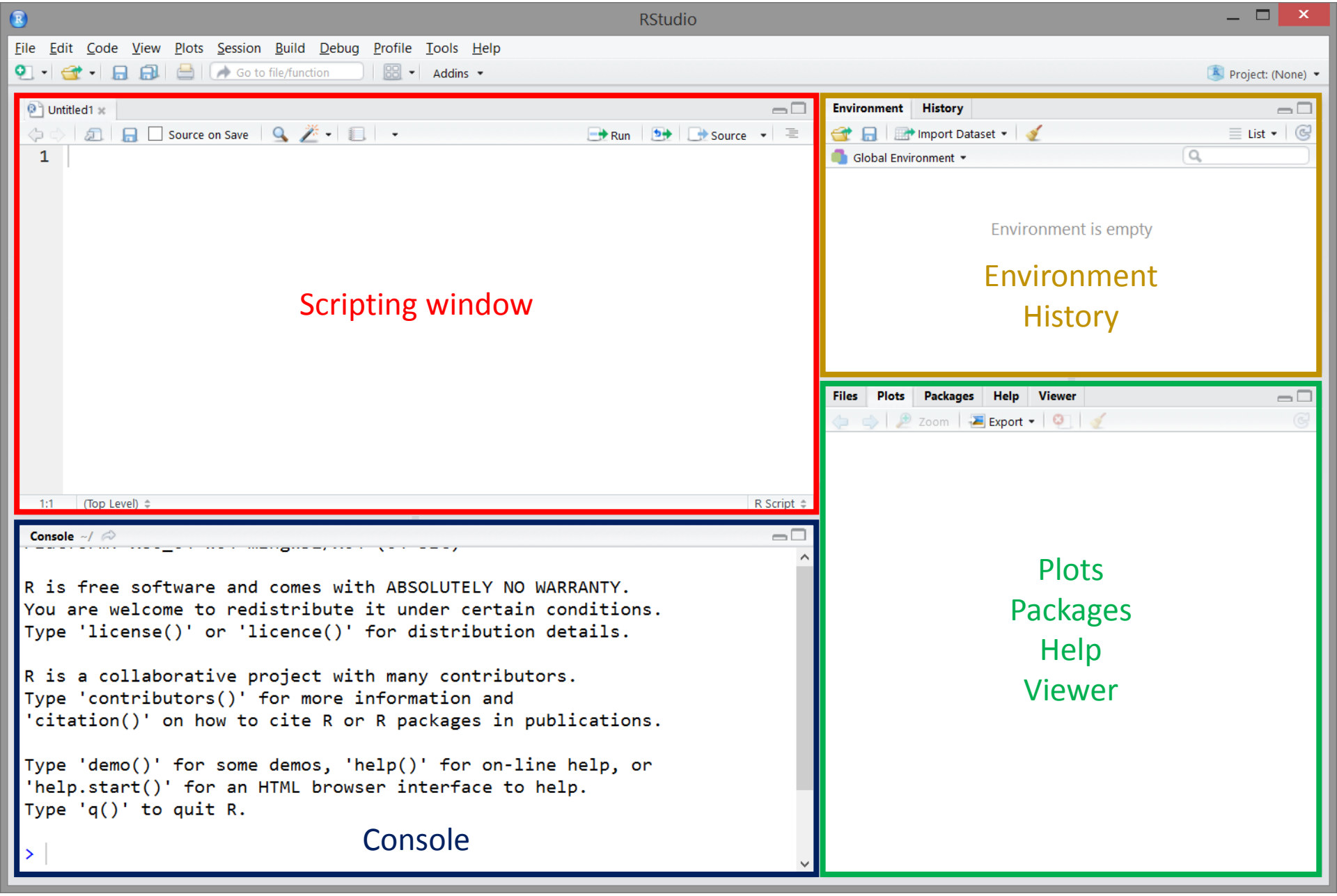
Introduction to R

- You should have two programs installed on computer
  - R
  - R Studio
- **Both** have to be installed to run **R Studio**
- We are going to use **R Studio**
  - More convenient to work with

R studio layout

Scripting window	Environment (stored objects) History
Console window	Plots Packages Help Viewer

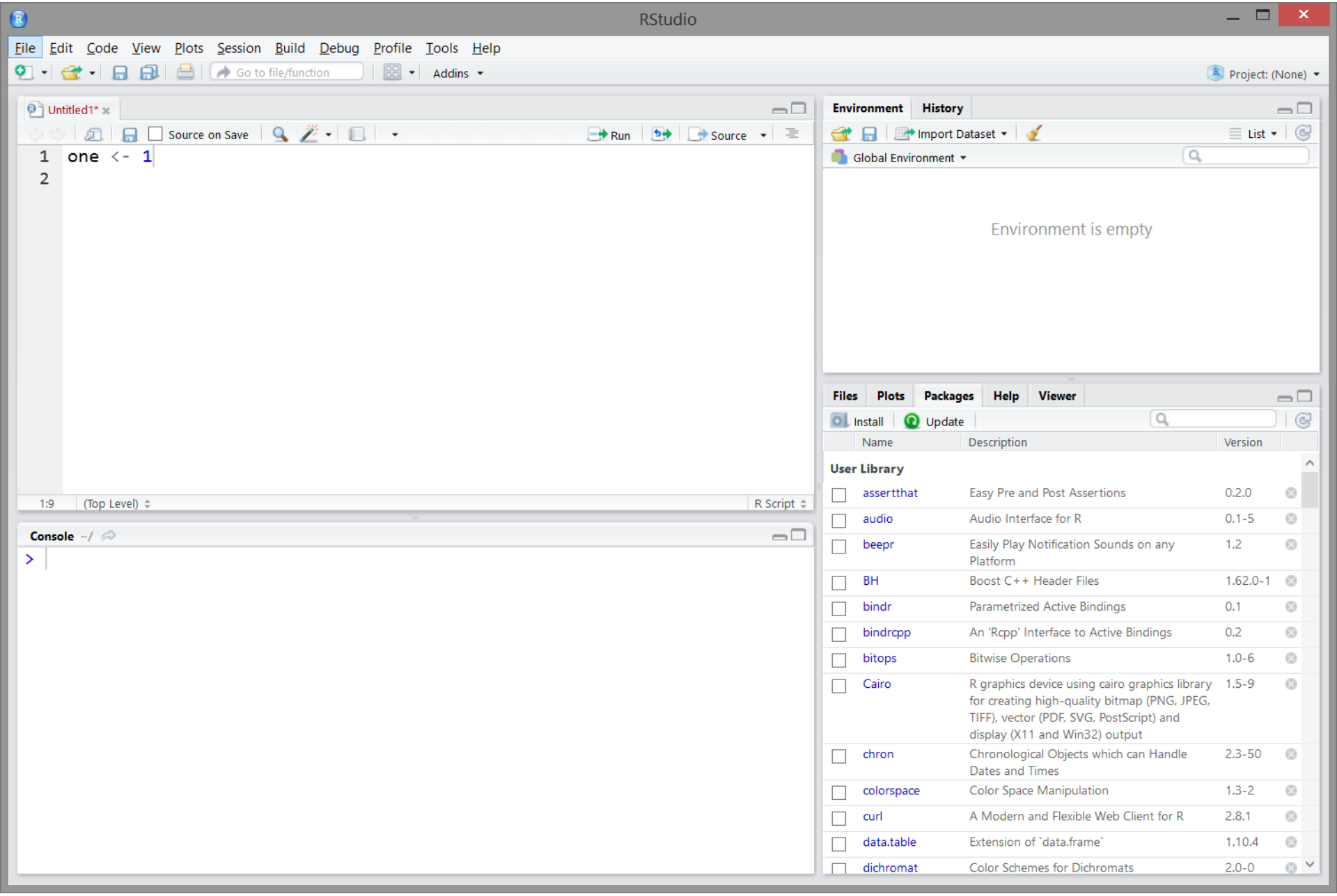

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Zoom
Export


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
Scripting window
1:1
(Top Level)
R Script
Console ~/
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Console
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Environment
History
Files
Plots
Packages
Help
Viewer
Zoom
Export
Plots
Packages
Help
Viewer

Object

- Object is a container which holds data, and can be manipulated with functions
- The most basic object is called **vector**
- There are other types of objects – **matrix**, **data frame**, **list**

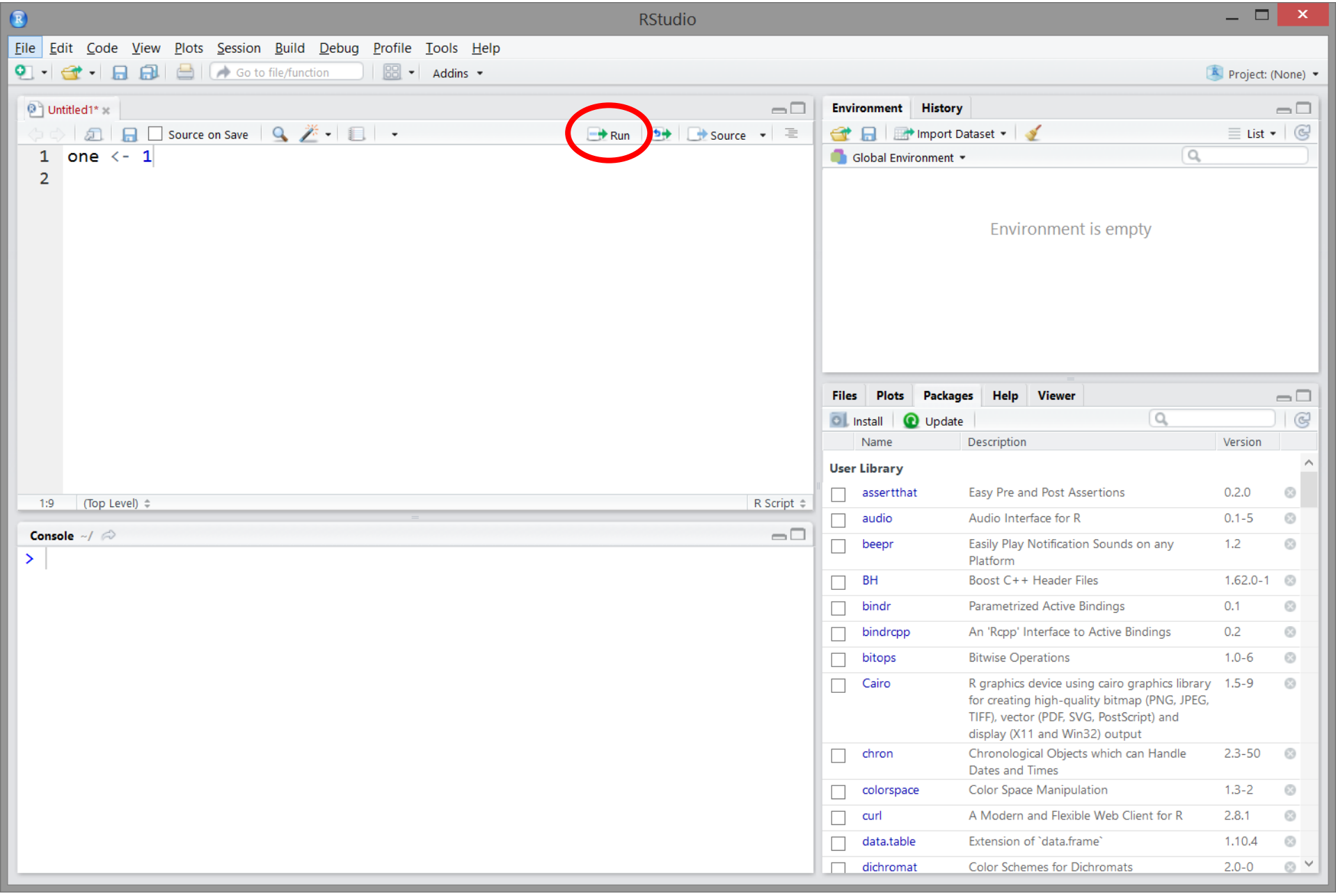
```
one <- 1
```

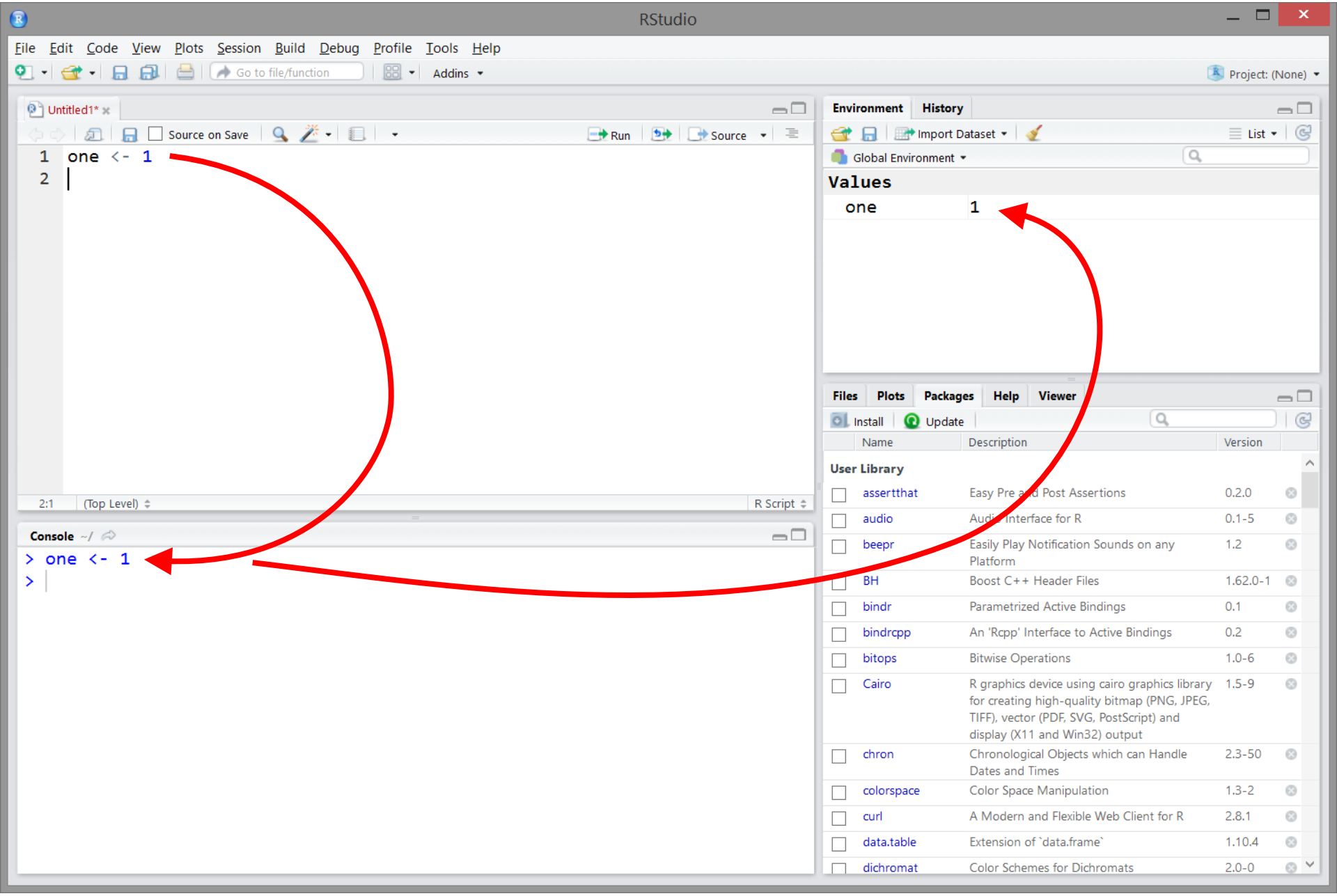

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
1:9
(Top Level)
R Script
Console ~/
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

Running commands from script

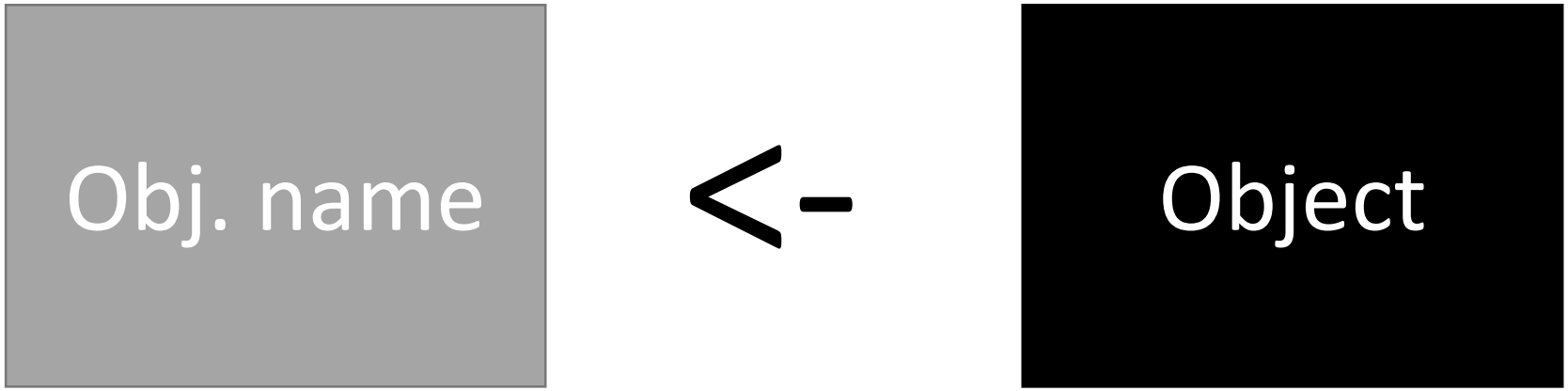
- Script is series of commands in one document (e.g. an application/program)
- Any region in script may be selected
- After selecting region, it may be executed
  - Hit button “Run” or use CTRL + R

- You may follow slides in the “R_crash_course.R” script


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
1:9
(Top Level)
R Script
Console ~/
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
2:1 (Top Level)
R Script
Console ~/
> one <- 1
>
Environment History
Import Dataset
List
Global Environment
Values
one 1
Files Plots Packages Help Viewer
Install Update
Name Description Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

Creating/storing objects



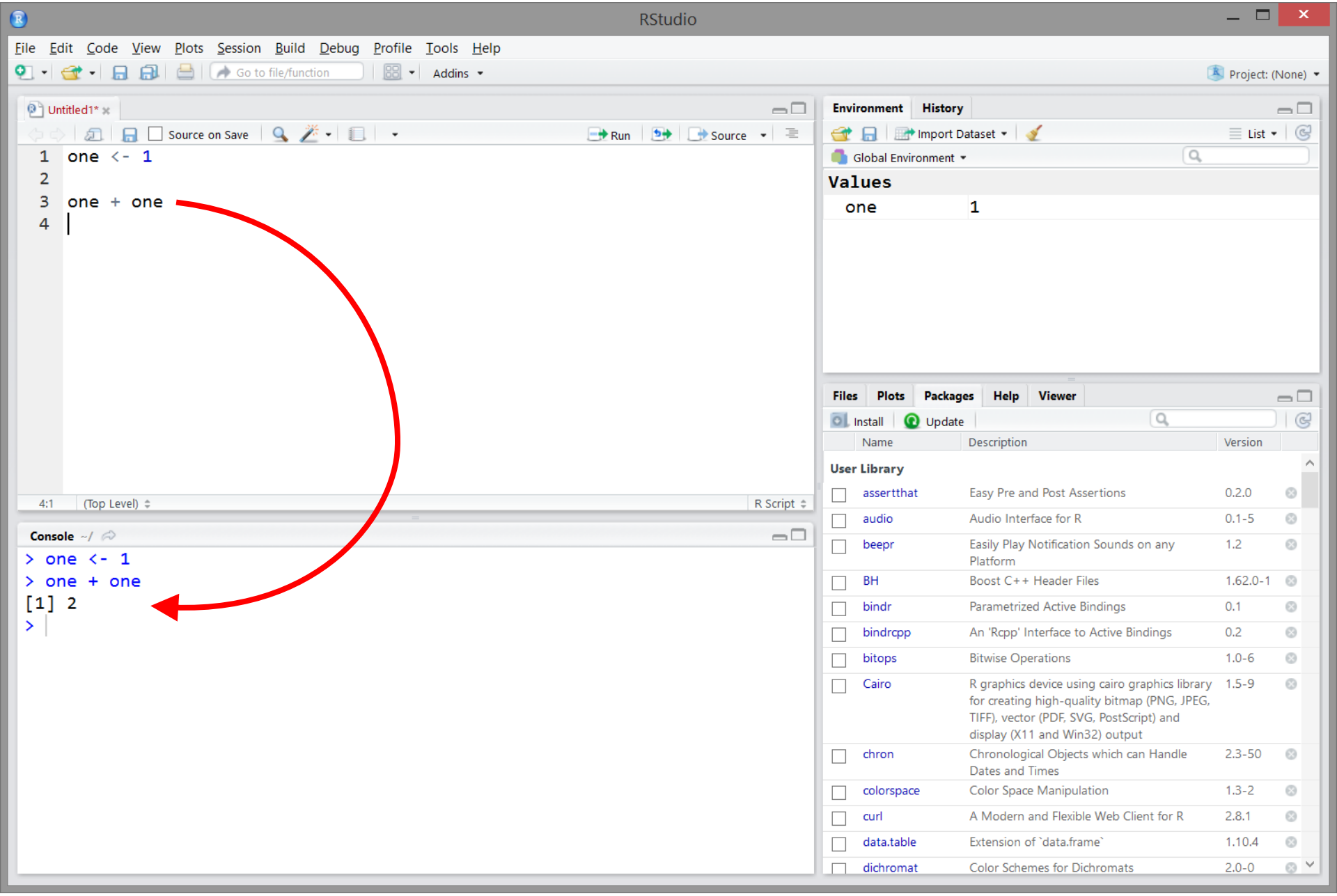
Object

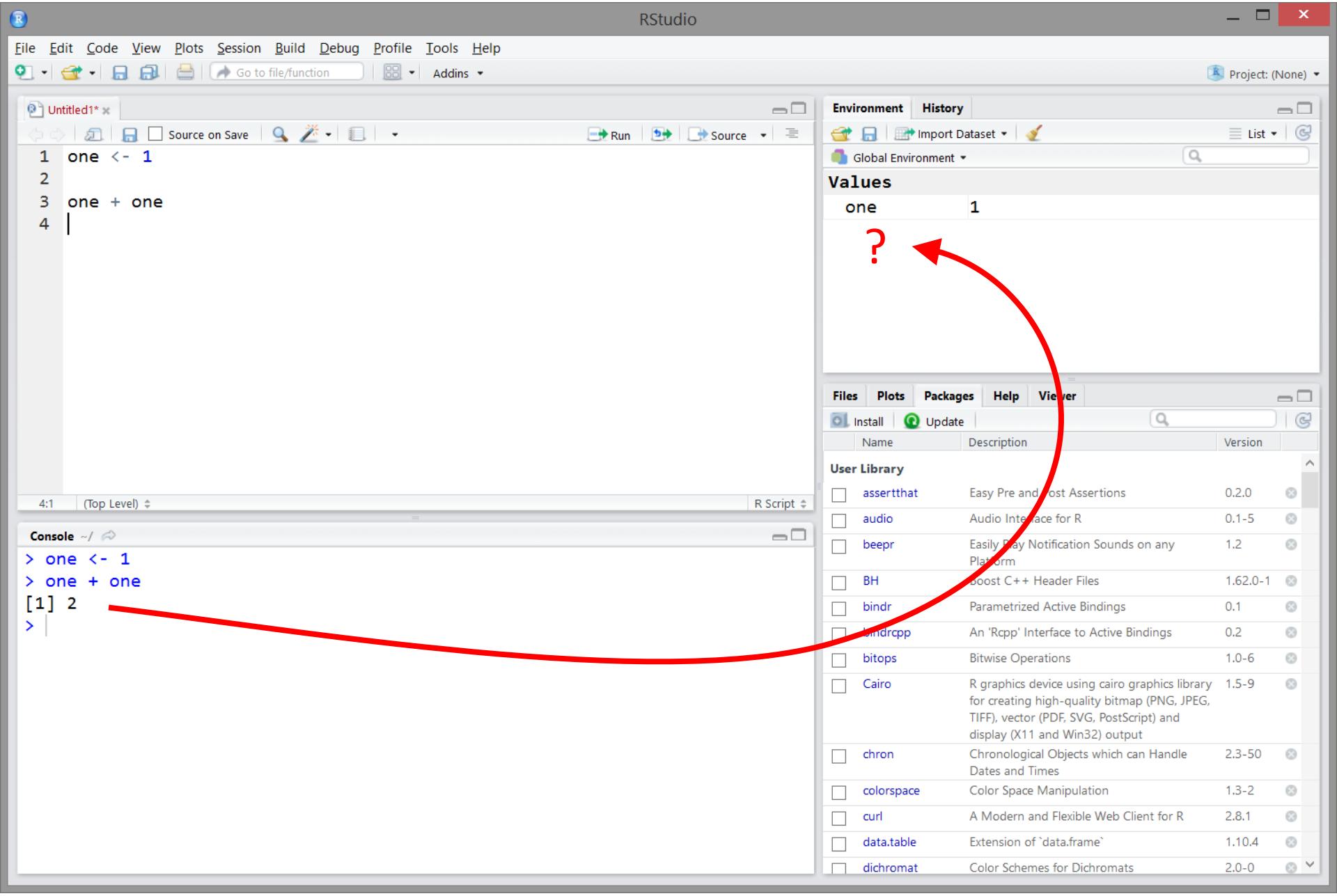
- Once objects exist, it may be used for operations

```
one <- 1

one + one

[1] 2
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
3 one + one
4
4:1
(Top Level)
R Script
Console ~/
> one <- 1
> one + one
[1] 2
>
Environment
History
Import Dataset
List
Global Environment
Values
one 1
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

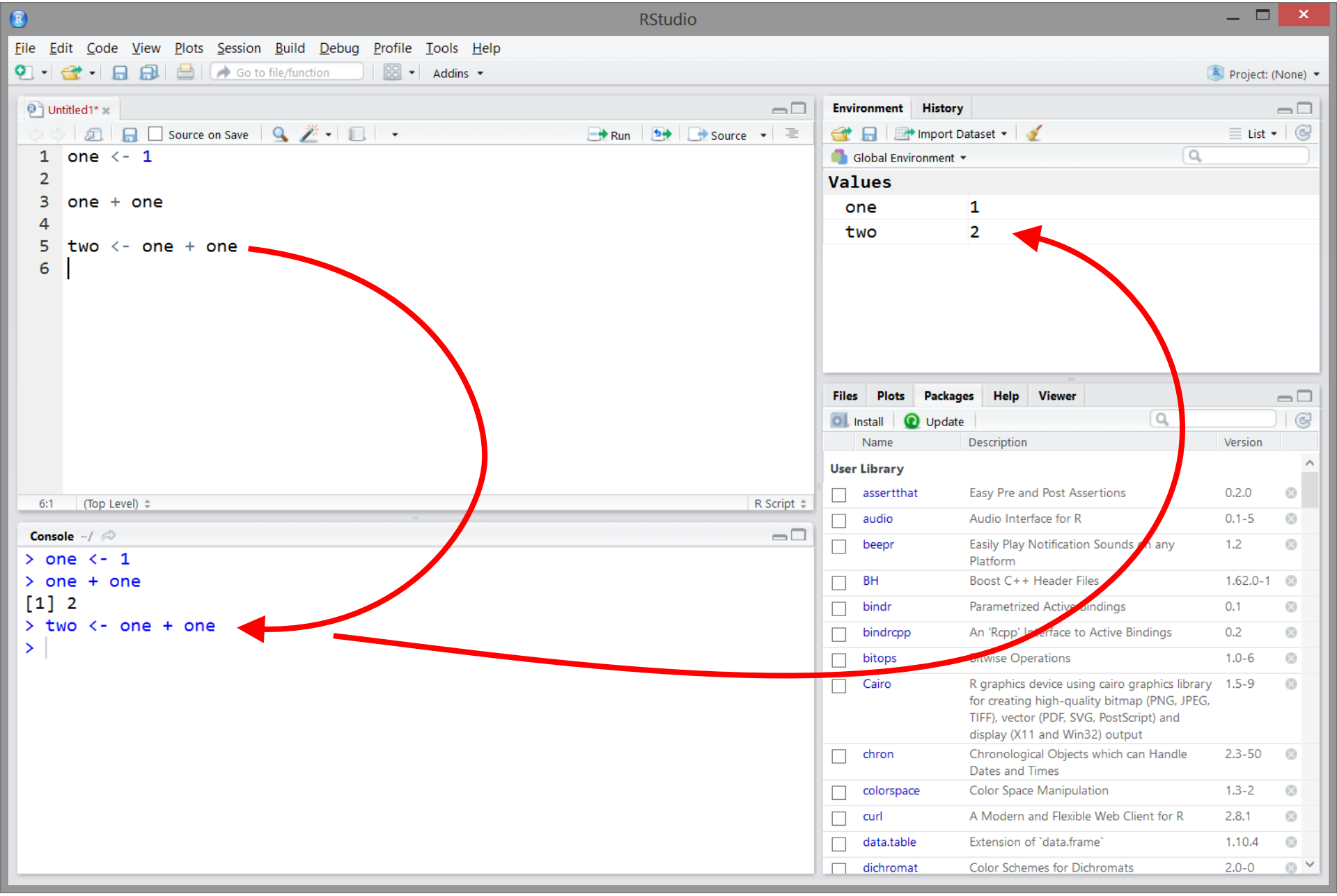

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
3 one + one
4
4:1
(Top Level)
R Script
Console ~/
> one <- 1
> one + one
[1] 2
>
Environment
History
Import Dataset
List
Global Environment
Values
one 1
?
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

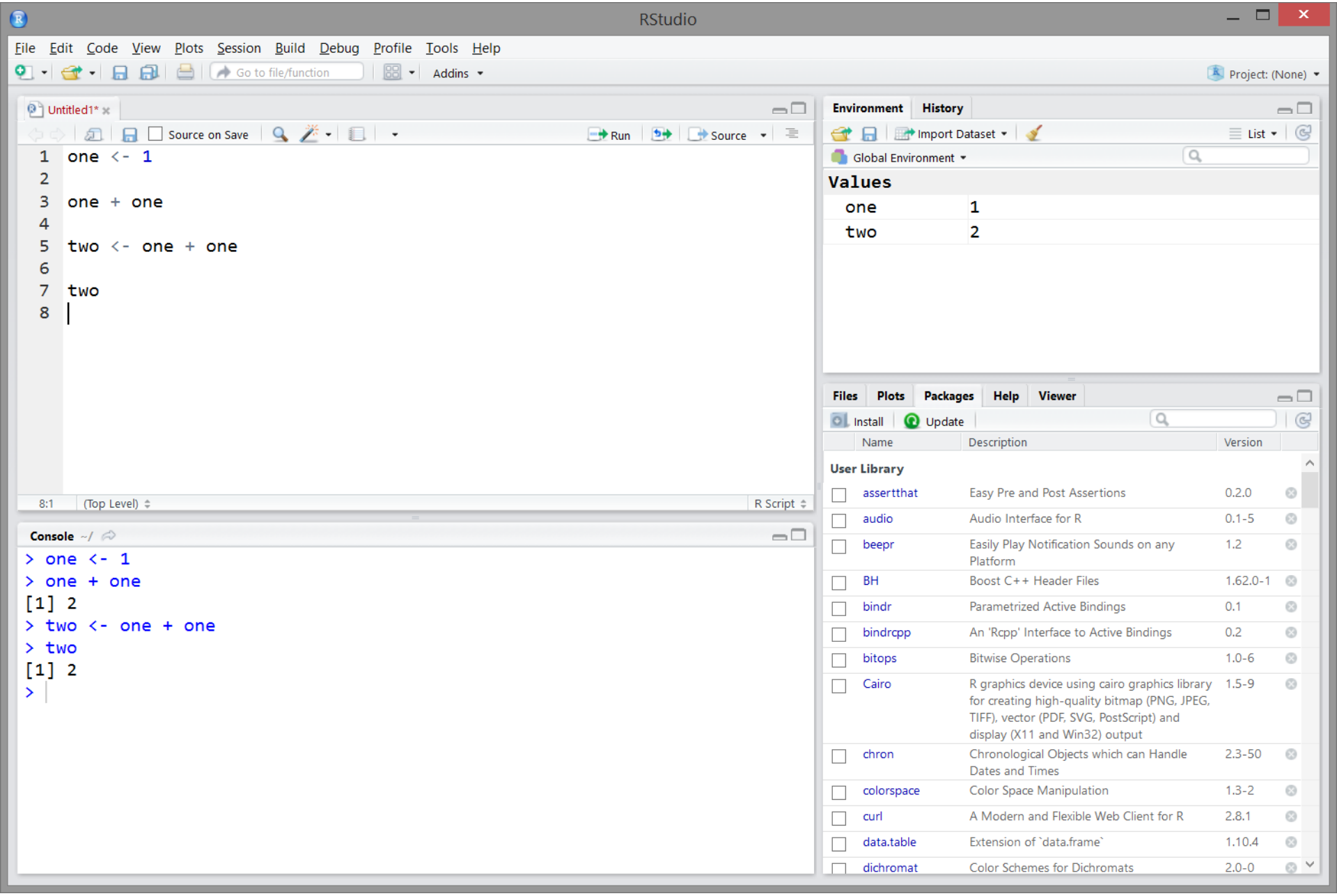
Object

- **Anything** may become an object
- New values must be stored as an object
  - **Conscious choice** to keep a result
  - **Object remains the same** unless overwritten
  - Must be **removed** by user as well

```
two <- one + one

two
[1] 2
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
3 one + one
4
5 two <- one + one
6
6:1
(Top Level)
R Script
Console ~/
> one <- 1
> one + one
[1] 2
> two <- one + one
>
Environment
History
Import Dataset
List
Global Environment
Values
one 1
two 2
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 one <- 1
2
3 one + one
4
5 two <- one + one
6
7 two
8
8:1
(Top Level)
R Script
Console ~/
> one <- 1
> one + one
[1] 2
> two <- one + one
> two
[1] 2
>
Environment
History
Import Dataset
List
Global Environment
Values
one 1
two 2
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

Object

- Any text input in R is always quoted
  - `"This is my piece of text"`
  - `"C:\\Users\\Lukas\\Documents"`
- Any text without quotes is considered a request for object
  - `myobject`
- Any text with parenthesis is considered a function
  - `sum()`

Functions

Functions

- Pre-defined **methods**

- To **create** an object with more than one element, function `c()` is used

```
onetofive <- c(1,3,5,4,2)
```

- Any object may be **manipulated** with a function

```
sort(onetofive)
[1] 1 2 3 4 5
```

Functions

- To extend functionality, functions have pre-defined **arguments**
  - **Arguments are further options** of functions
  - Some functions have many arguments, some none
- To keep function result, it must be stored in the environment as an object

```
sort(onetofive)
[1] 1 2 3 4 5

sort(onetofive, decreasing = TRUE)
[1] 5 4 3 2 1

onetofive <- sort(onetofive, decreasing = TRUE)
```

Functions

- `fun()` – parentheses indicate a function

```
sqrt(9)
[1] 3
```

- Structure is `function(arg1, arg2, …)`

```
sample(0:100, 10, rep = FALSE)
[1] 48 50 37 94 42 39 21 19 63 95
```

Functions

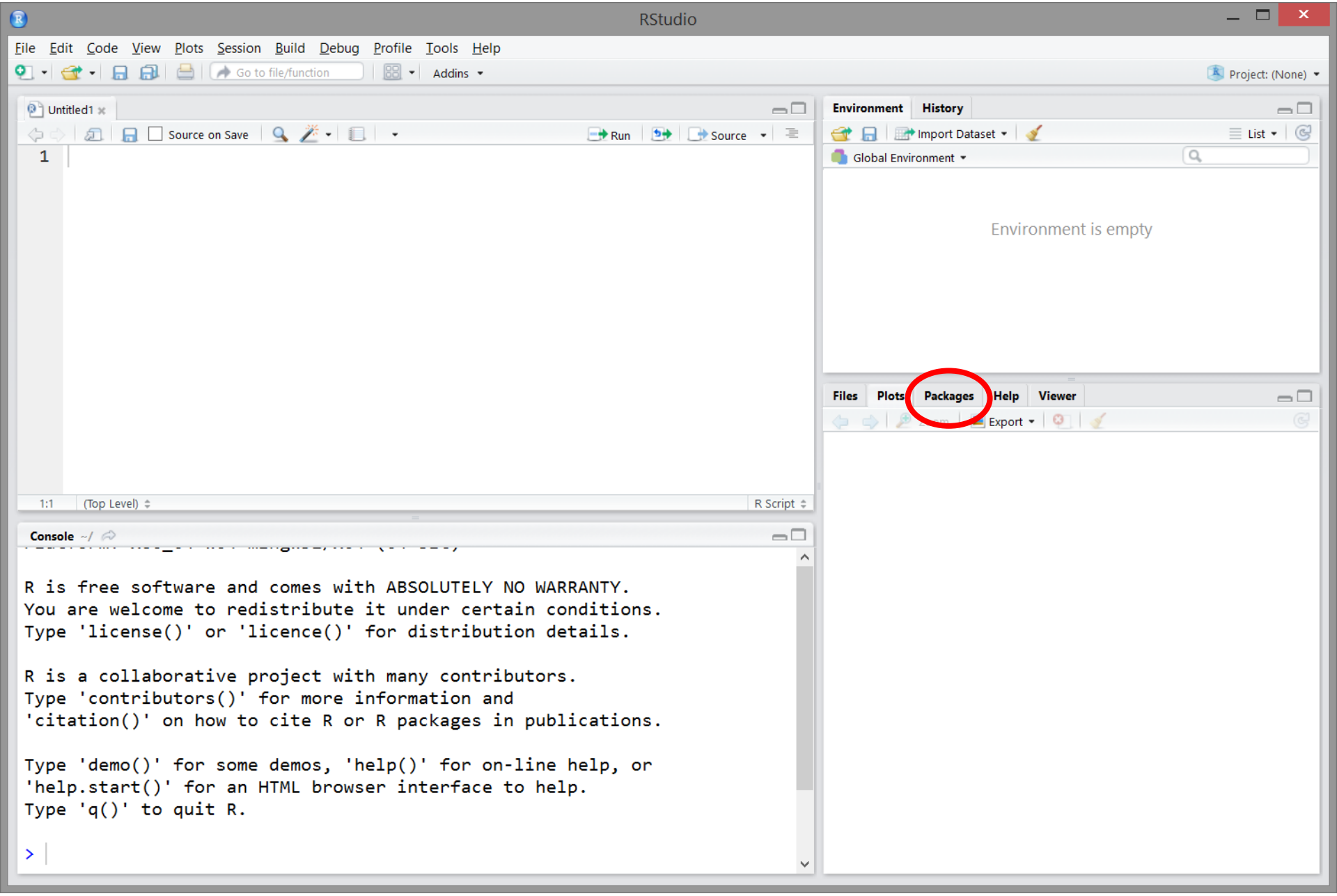
- Arguments usually require input format
  - Boolean input – `TRUE` or `FALSE`
  - Name of object – `onetofive`
  - Text value – `"linear"`
- Format of each argument may be found in **help page**
  - Just **add question mark** in front of the function name

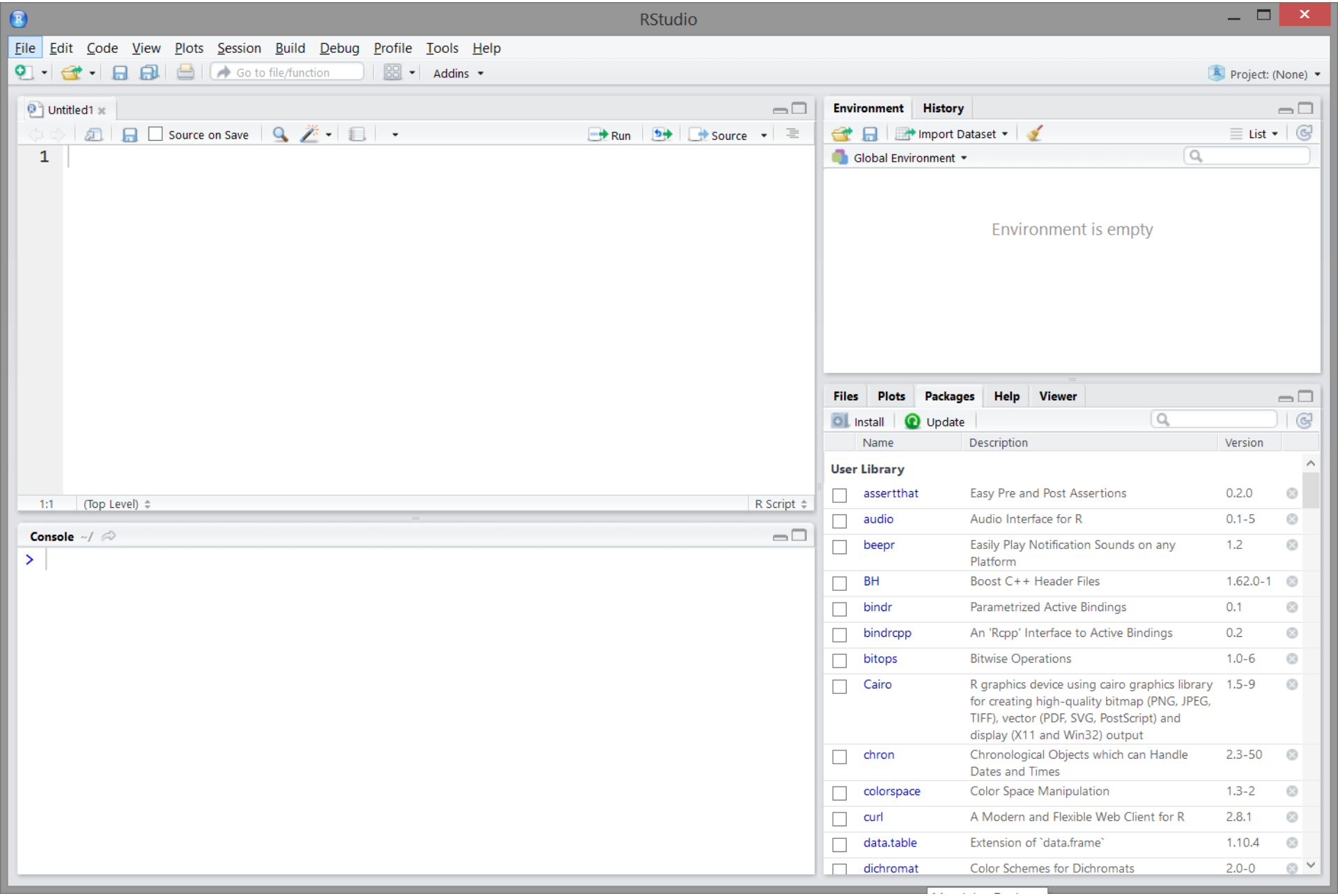
```
?sample()
```

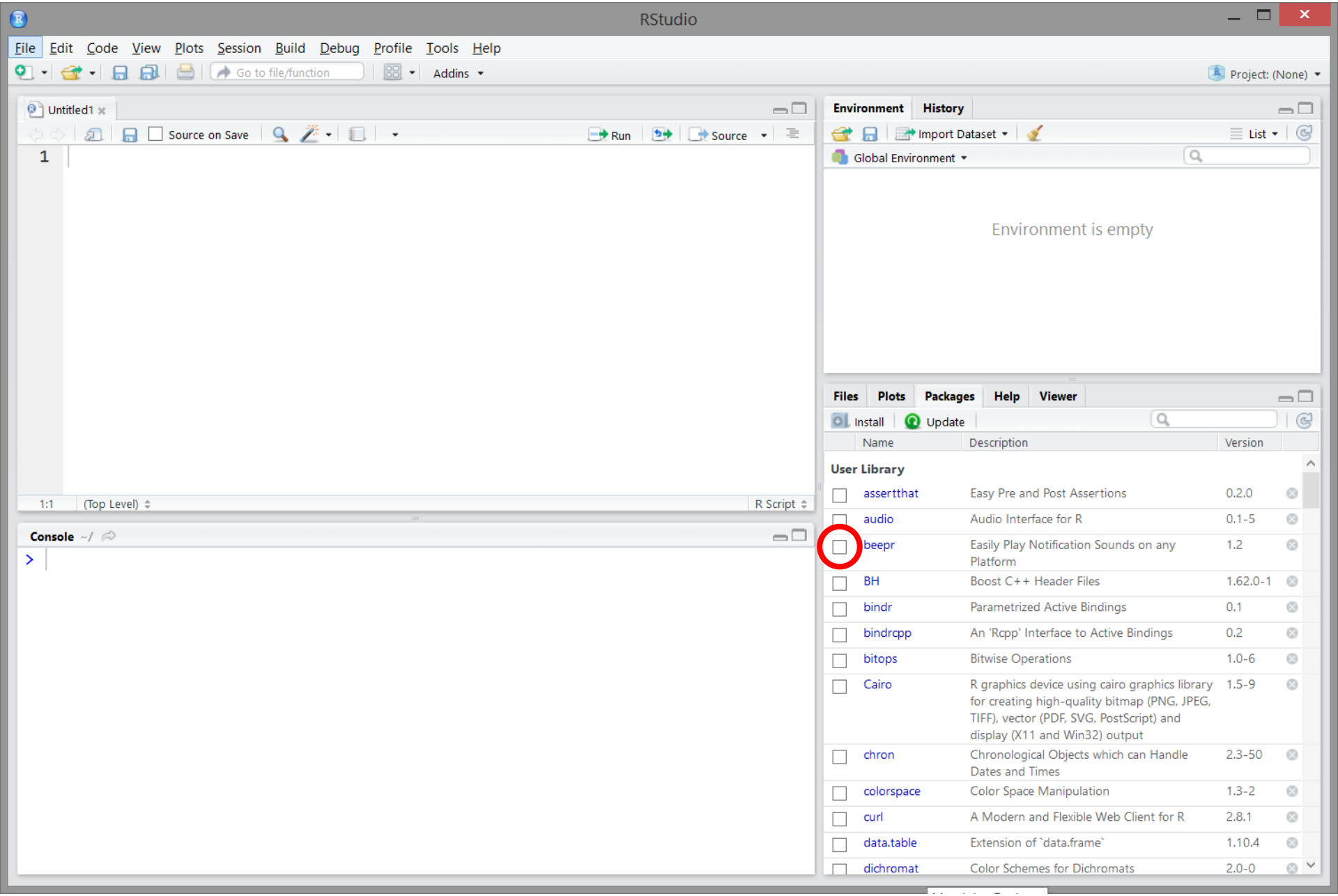
Libraries/packages

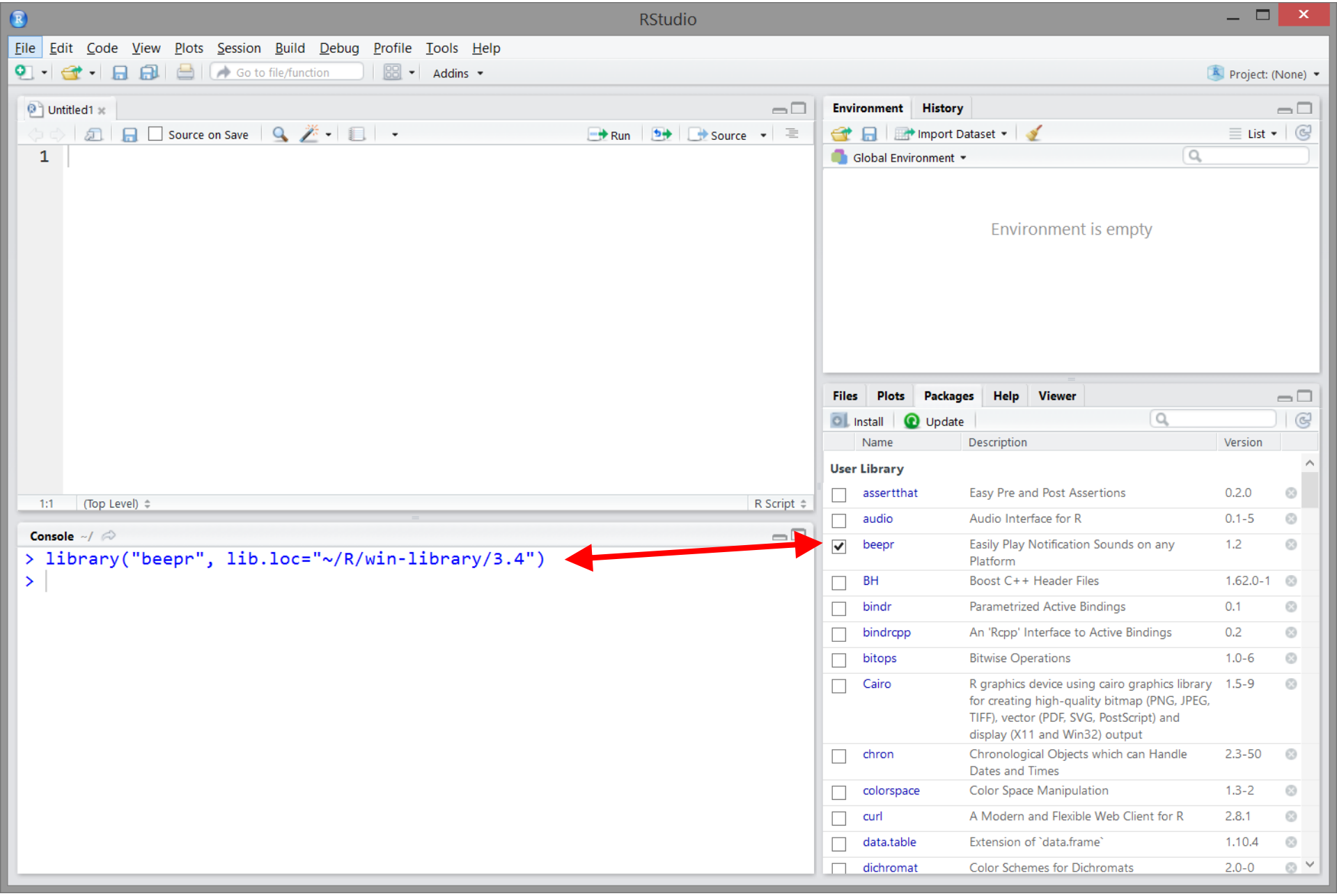
Library

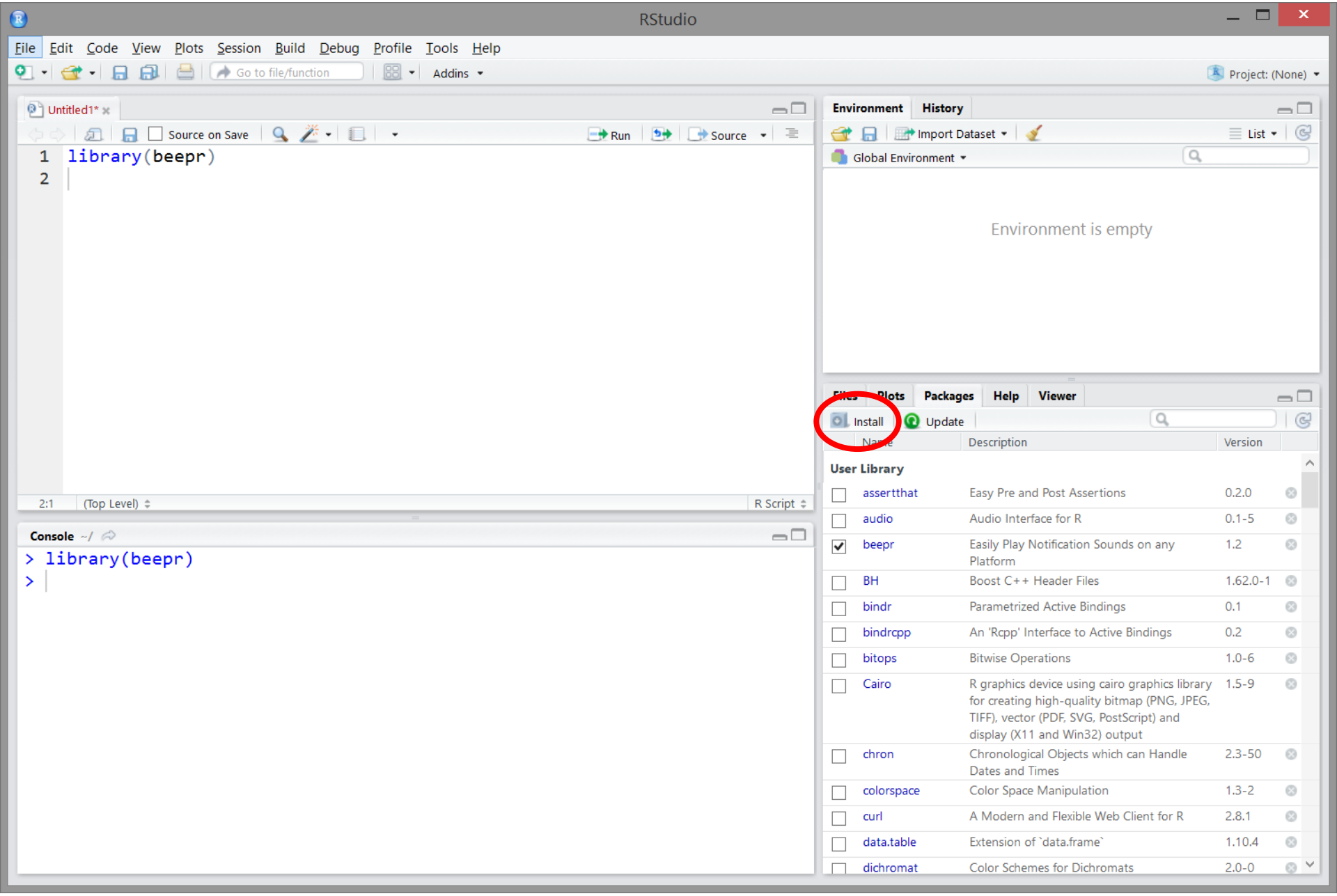
- Libraries are like mobile applications – allow to load functions according to problem at hand
  - Load, install and unload either using R Studio or using a function in script
  - Libraries download and install automatically from R repository on the web
- **Every time you start** fresh R session, you have to **reload** all libraries

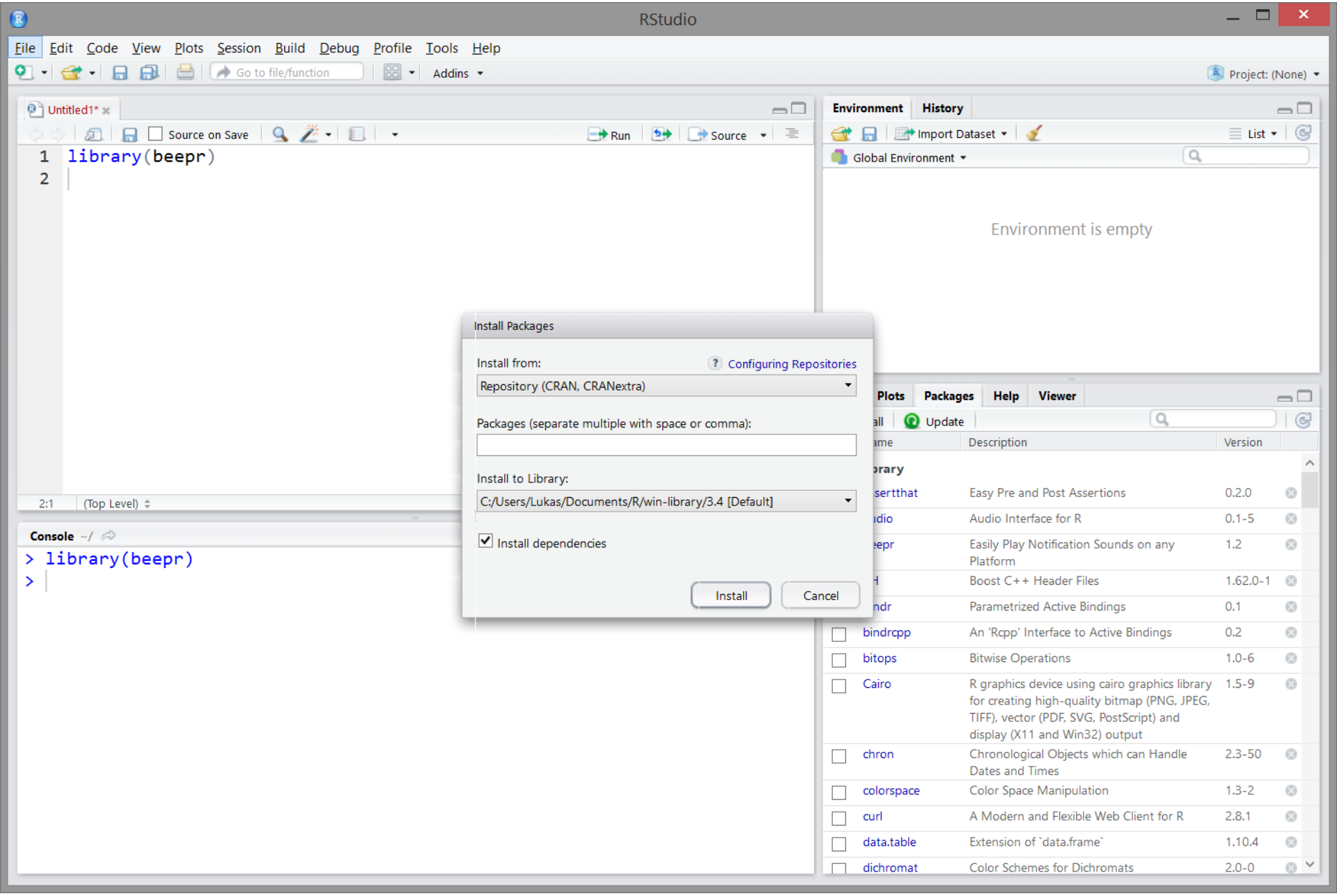

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Export

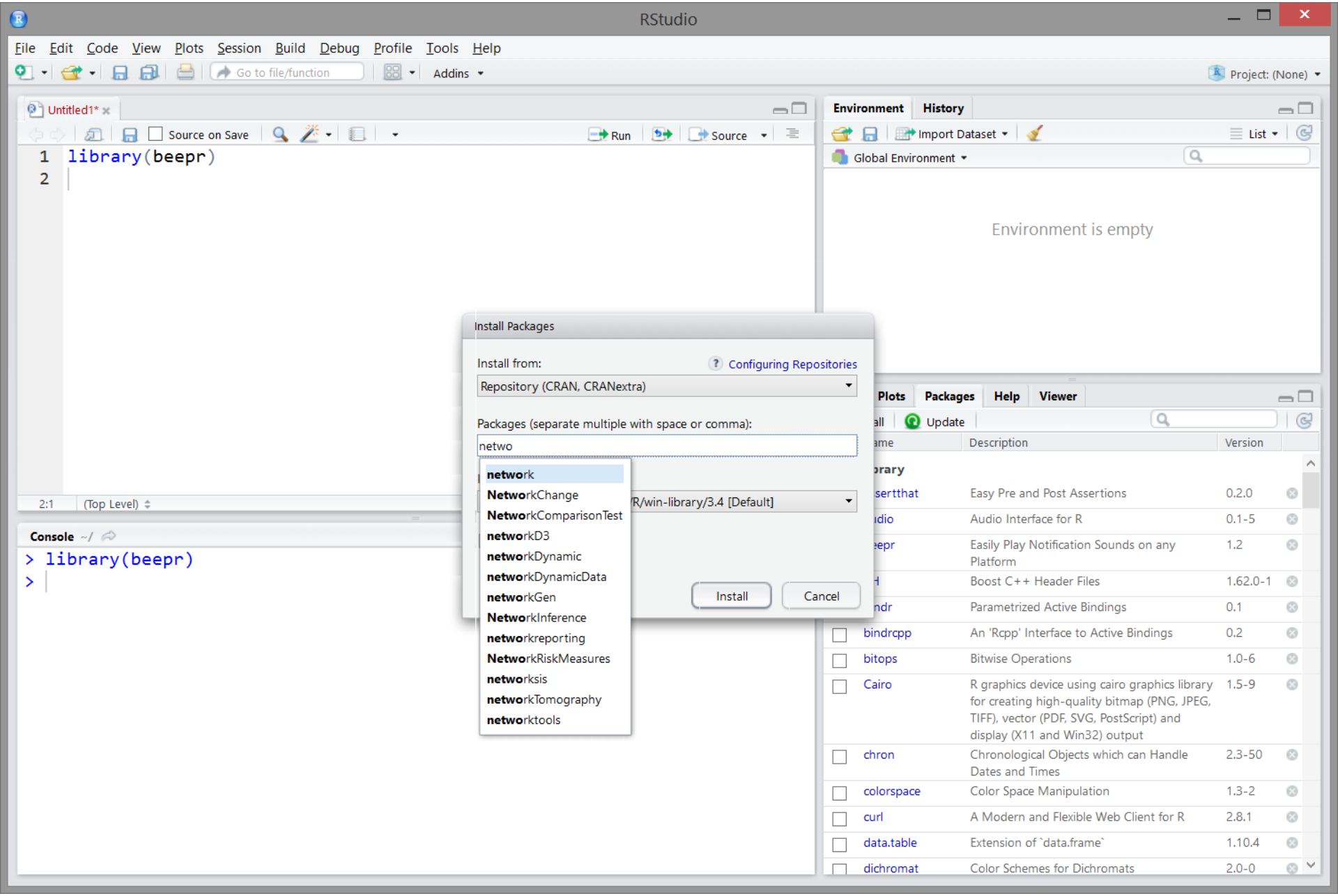

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

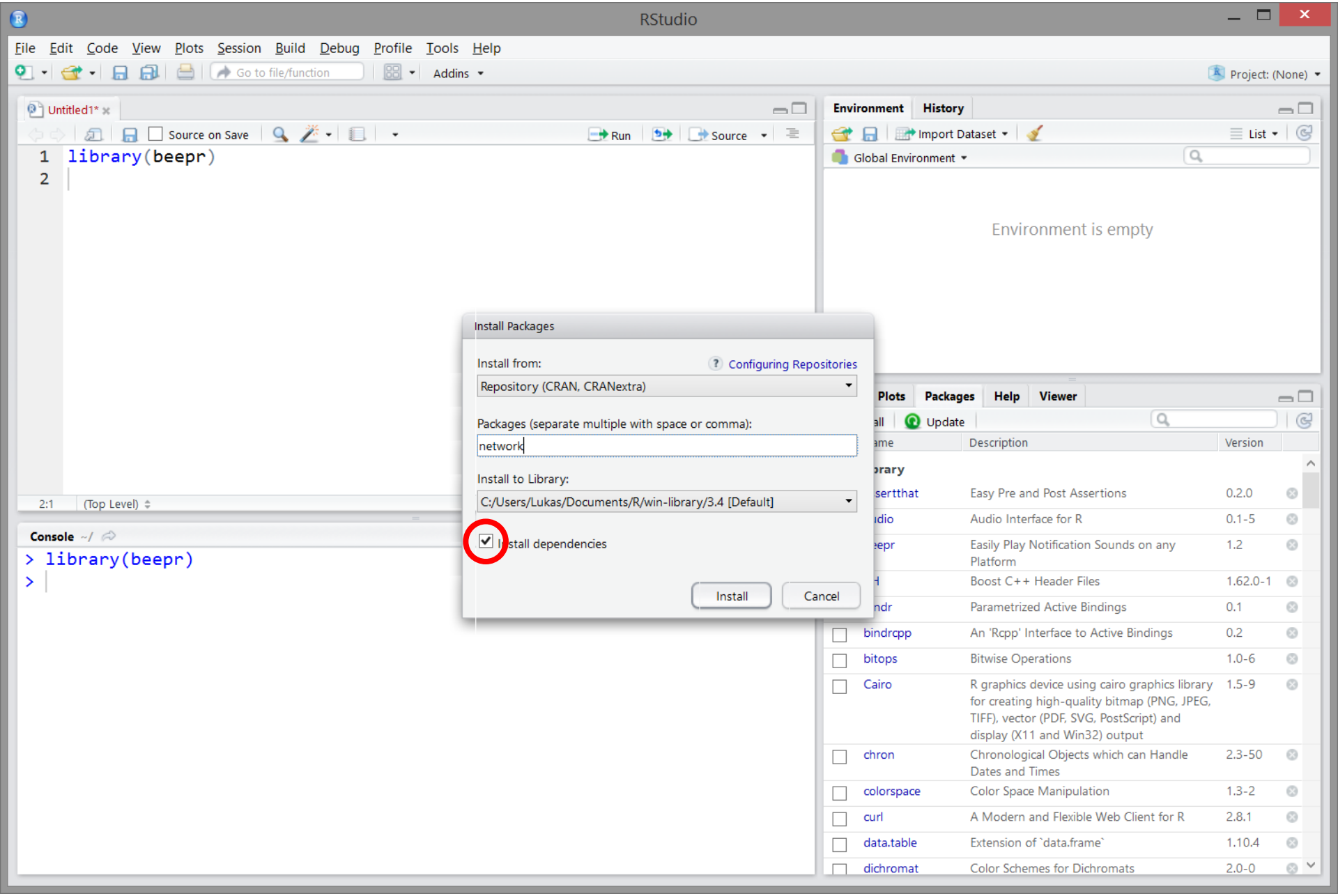

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

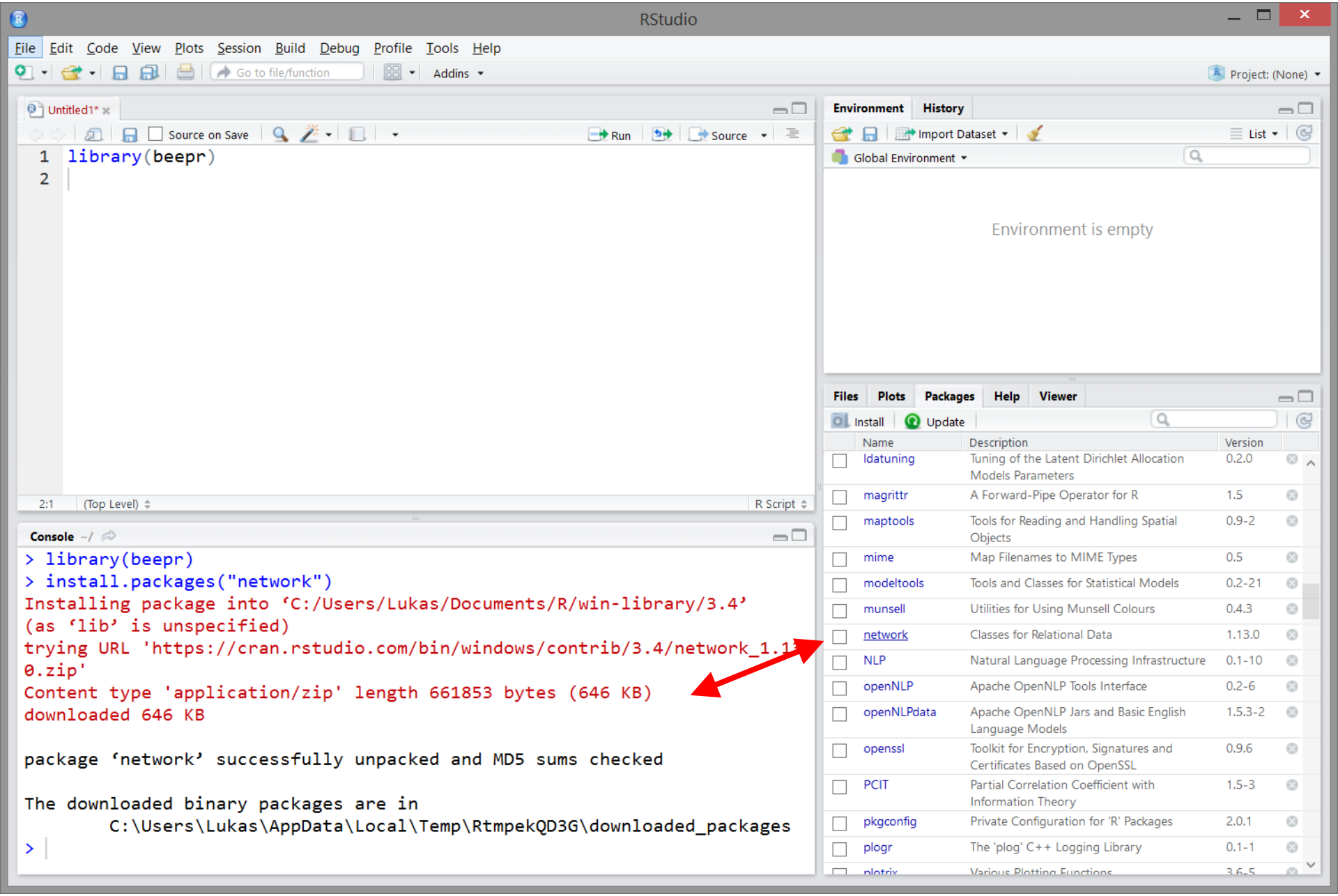

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1
Source on Save
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
> library("beepr", lib.loc="~/R/win-library/3.4")
>
Environment
History
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
Install
Update
Name
Description
Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

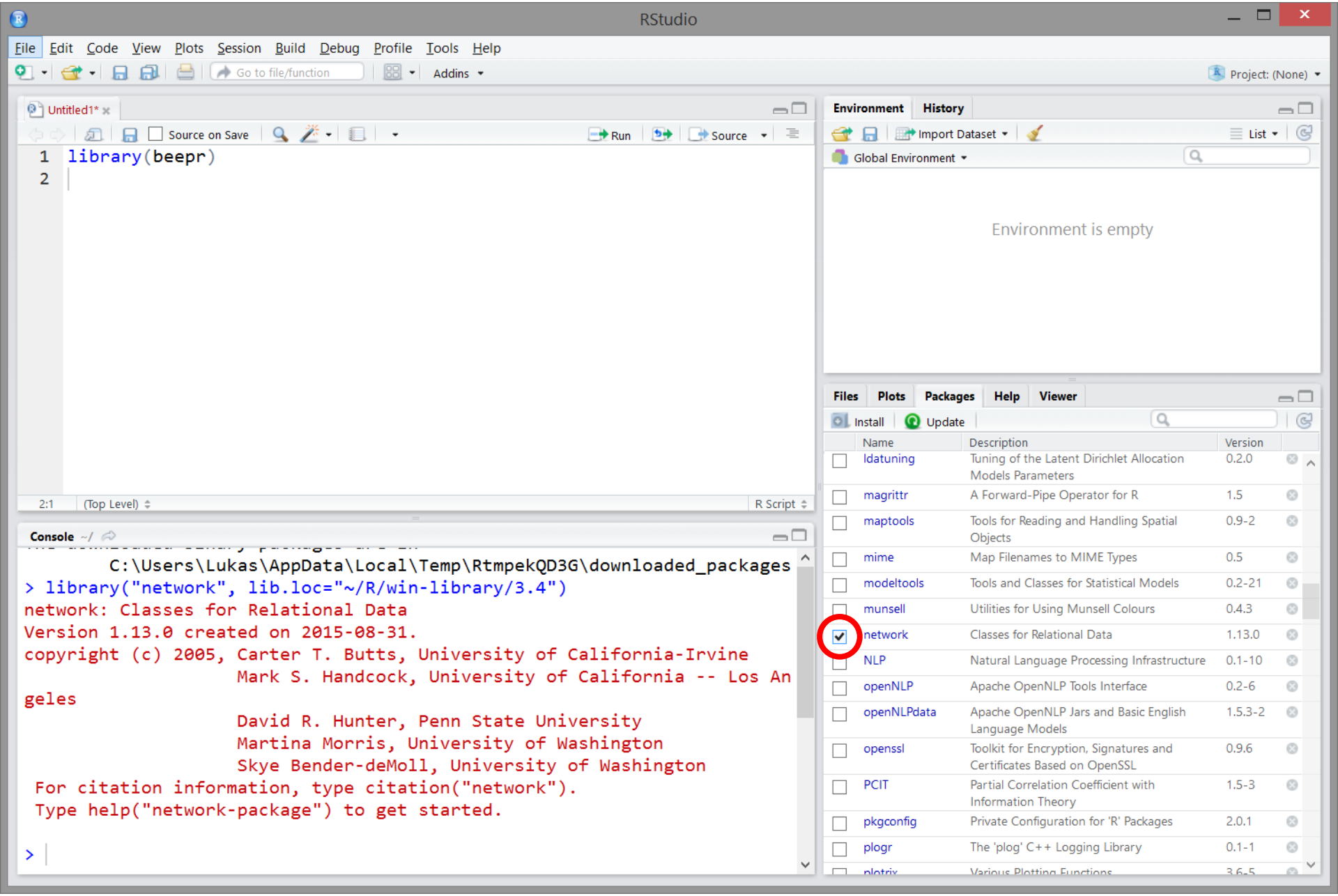

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 library(beepr)
2
2:1 (Top Level)
R Script
Console ~/
> library(beepr)
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name Description Version
User Library
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0

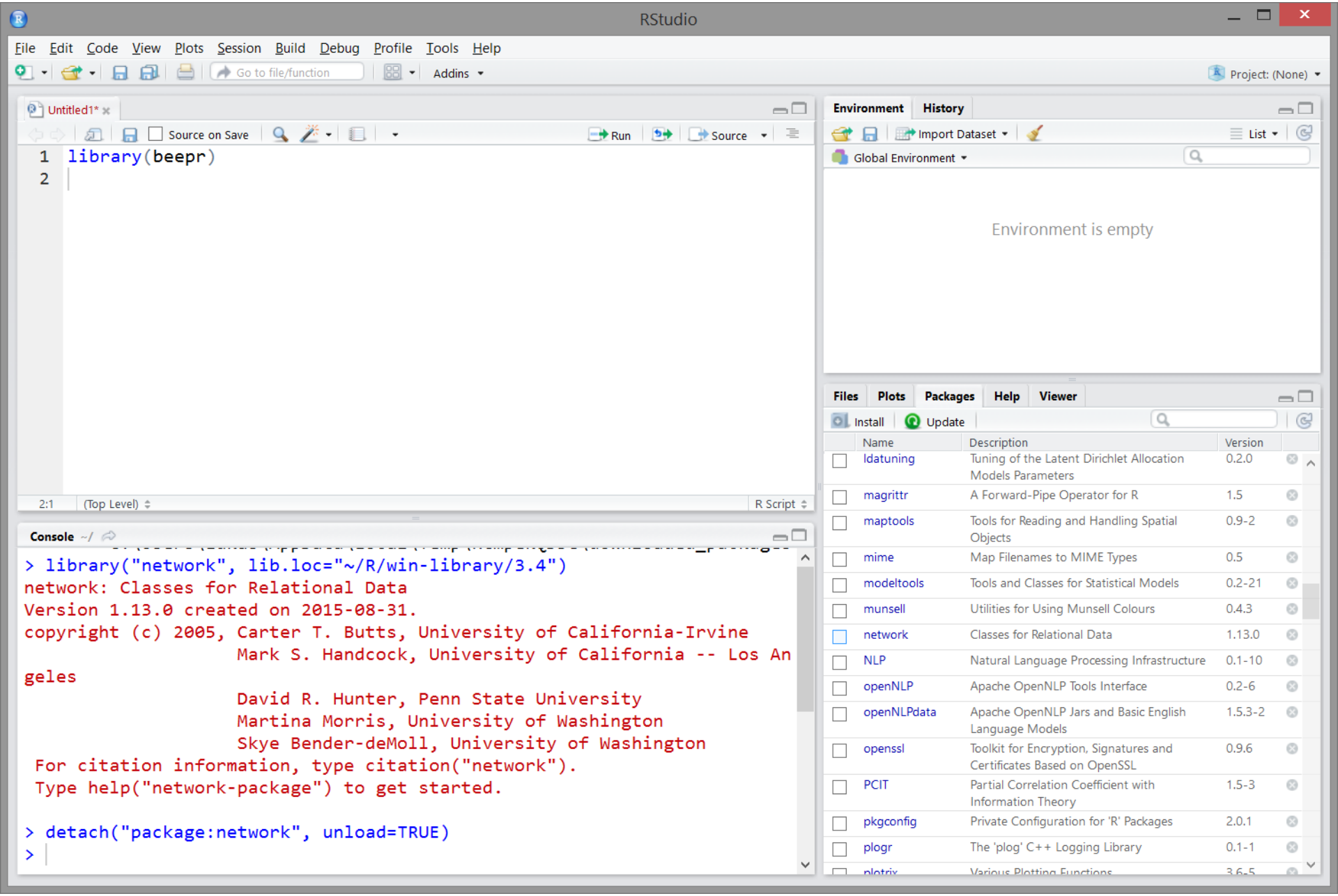

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 library(beepr)
2
2:1 (Top Level)
Console ~/
> library(beepr)
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Install Packages
Install from:
Configuring Repositories
Repository (CRAN, CRANextra)
Packages (separate multiple with space or comma):
Install to Library:
C:/Users/Lukas/Documents/R/win-library/3.4 [Default]
Install dependencies
Install
Cancel
Plots Packages Help Viewer
Update
Description
Version
Easy Pre and Post Assertions 0.2.0
Audio Interface for R 0.1-5
Easily Play Notification Sounds on any Platform 1.2
Boost C++ Header Files 1.62.0-1
Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 library(beepr)
2
2:1 (Top Level)
Console ~/
> library(beepr)
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Install Packages
Install from:
Configuring Repositories
Repository (CRAN, CRANextra)
Packages (separate multiple with space or comma):
netwo
network
NetworkChange
NetworkComparisonTest
networkD3
networkDynamic
networkDynamicData
networkGen
NetworkInference
networkreporting
NetworkRiskMeasures
networksis
networkTomography
networktools
R/win-library/3.4 [Default]
Install
Cancel
Plots Packages Help Viewer
Update
Description
Version
Easy Pre and Post Assertions 0.2.0
Audio Interface for R 0.1-5
Easily Play Notification Sounds on any Platform 1.2
Boost C++ Header Files 1.62.0-1
Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 library(beepr)
2
2:1 (Top Level)
Console ~/
> library(beepr)
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Install Packages
Install from:
Configuring Repositories
Repository (CRAN, CRANextra)
Packages (separate multiple with space or comma):
network
Install to Library:
C:/Users/Lukas/Documents/R/win-library/3.4 [Default]
Install dependencies
Install
Cancel
Plots Packages Help Viewer
Update
Description
Version
Easy Pre and Post Assertions 0.2.0
Audio Interface for R 0.1-5
Easily Play Notification Sounds on any Platform 1.2
Boost C++ Header Files 1.62.0-1
Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9
chron Chronological Objects which can Handle Dates and Times 2.3-50
colorspace Color Space Manipulation 1.3-2
curl A Modern and Flexible Web Client for R 2.8.1
data.table Extension of `data.frame` 1.10.4
dichromat Color Schemes for Dichromats 2.0-0


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 library(beepr)
2
2:1 (Top Level)
R Script
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name Description Version
ldatuning Tuning of the Latent Dirichlet Allocation Models Parameters 0.2.0
magrittr A Forward-Pipe Operator for R 1.5
maptools Tools for Reading and Handling Spatial Objects 0.9-2
mime Map Filenames to MIME Types 0.5
modeltools Tools and Classes for Statistical Models 0.2-21
munsell Utilities for Using Munsell Colours 0.4.3
network Classes for Relational Data 1.13.0
NLP Natural Language Processing Infrastructure 0.1-10
openNLP Apache OpenNLP Tools Interface 0.2-6
openNLPdata Apache OpenNLP Jars and Basic English Language Models 1.5.3-2
openssl Toolkit for Encryption, Signatures and Certificates Based on OpenSSL 0.9.6
PCIT Partial Correlation Coefficient with Information Theory 1.5-3
pkgconfig Private Configuration for 'R' Packages 2.0.1
plogr The 'plog' C++ Logging Library 0.1-1
Console ~/
> library(beepr)
> install.packages("network")
Installing package into 'C:/Users/Lukas/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/network_1.13.0.zip'
Content type 'application/zip' length 661853 bytes (646 KB)
downloaded 646 KB
package 'network' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
C:\Users\Lukas\AppData\Local\Temp\RtmpekQD3G\downloaded_packages
>


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 library(beepr)
2
2:1 (Top Level)
R Script
Console ~/
C:\Users\Lukas\AppData\Local\Temp\RtmpekQD3G\downloaded_packages
> library("network", lib.loc="~/R/win-library/3.4")
network: Classes for Relational Data
Version 1.13.0 created on 2015-08-31.
copyright (c) 2005, Carter T. Butts, University of California-Irvine
Mark S. Handcock, University of California -- Los Angeles
David R. Hunter, Penn State University
Martina Morris, University of Washington
Skye Bender-deMoll, University of Washington
For citation information, type citation("network").
Type help("network-package") to get started.
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name Description Version
ldatuning Tuning of the Latent Dirichlet Allocation Models Parameters 0.2.0
magrittr A Forward-Pipe Operator for R 1.5
maptools Tools for Reading and Handling Spatial Objects 0.9-2
mime Map Filenames to MIME Types 0.5
modeltools Tools and Classes for Statistical Models 0.2-21
munsell Utilities for Using Munsell Colours 0.4.3
network Classes for Relational Data 1.13.0
NLP Natural Language Processing Infrastructure 0.1-10
openNLP Apache OpenNLP Tools Interface 0.2-6
openNLPdata Apache OpenNLP Jars and Basic English Language Models 1.5.3-2
openssl Toolkit for Encryption, Signatures and Certificates Based on OpenSSL 0.9.6
PCIT Partial Correlation Coefficient with Information Theory 1.5-3
pkgconfig Private Configuration for 'R' Packages 2.0.1
plogr The 'plog' C++ Logging Library 0.1-1

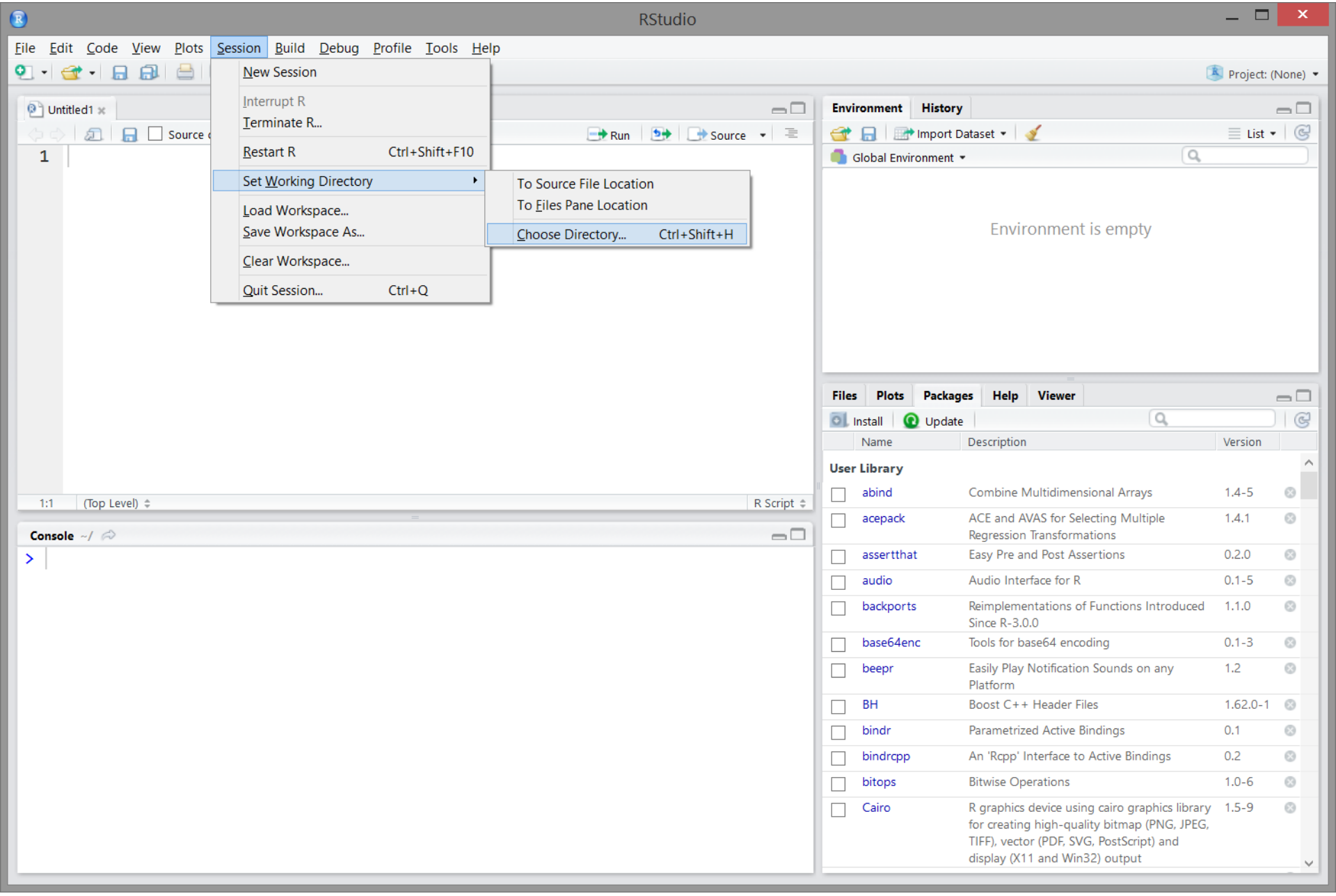

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
Source on Save
Run
Source
1 library(beepr)
2
2:1 (Top Level)
R Script
Console ~/
> library("network", lib.loc="~/R/win-library/3.4")
network: Classes for Relational Data
Version 1.13.0 created on 2015-08-31.
copyright (c) 2005, Carter T. Butts, University of California-Irvine
Mark S. Handcock, University of California -- Los An
geles
David R. Hunter, Penn State University
Martina Morris, University of Washington
Skye Bender-deMoll, University of Washington
For citation information, type citation("network").
Type help("network-package") to get started.
> detach("package:network", unload=TRUE)
>
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name Description Version
ldatuning Tuning of the Latent Dirichlet Allocation Models Parameters 0.2.0
magrittr A Forward-Pipe Operator for R 1.5
maptools Tools for Reading and Handling Spatial Objects 0.9-2
mime Map Filenames to MIME Types 0.5
modeltools Tools and Classes for Statistical Models 0.2-21
munsell Utilities for Using Munsell Colours 0.4.3
network Classes for Relational Data 1.13.0
NLP Natural Language Processing Infrastructure 0.1-10
openNLP Apache OpenNLP Tools Interface 0.2-6
openNLPdata Apache OpenNLP Jars and Basic English Language Models 1.5.3-2
openssl Toolkit for Encryption, Signatures and Certificates Based on OpenSSL 0.9.6
PCIT Partial Correlation Coefficient with Information Theory 1.5-3
pkgconfig Private Configuration for 'R' Packages 2.0.1
plogr The 'plog' C++ Logging Library 0.1-1

Working directory

- Folder, where everything is taking place – enough to set once
- Makes data import and export easier
- Function `setwd()`
- Does not accept single backslash in Win path
  - Replace backslash \ with forwardslash / or double backslash \\

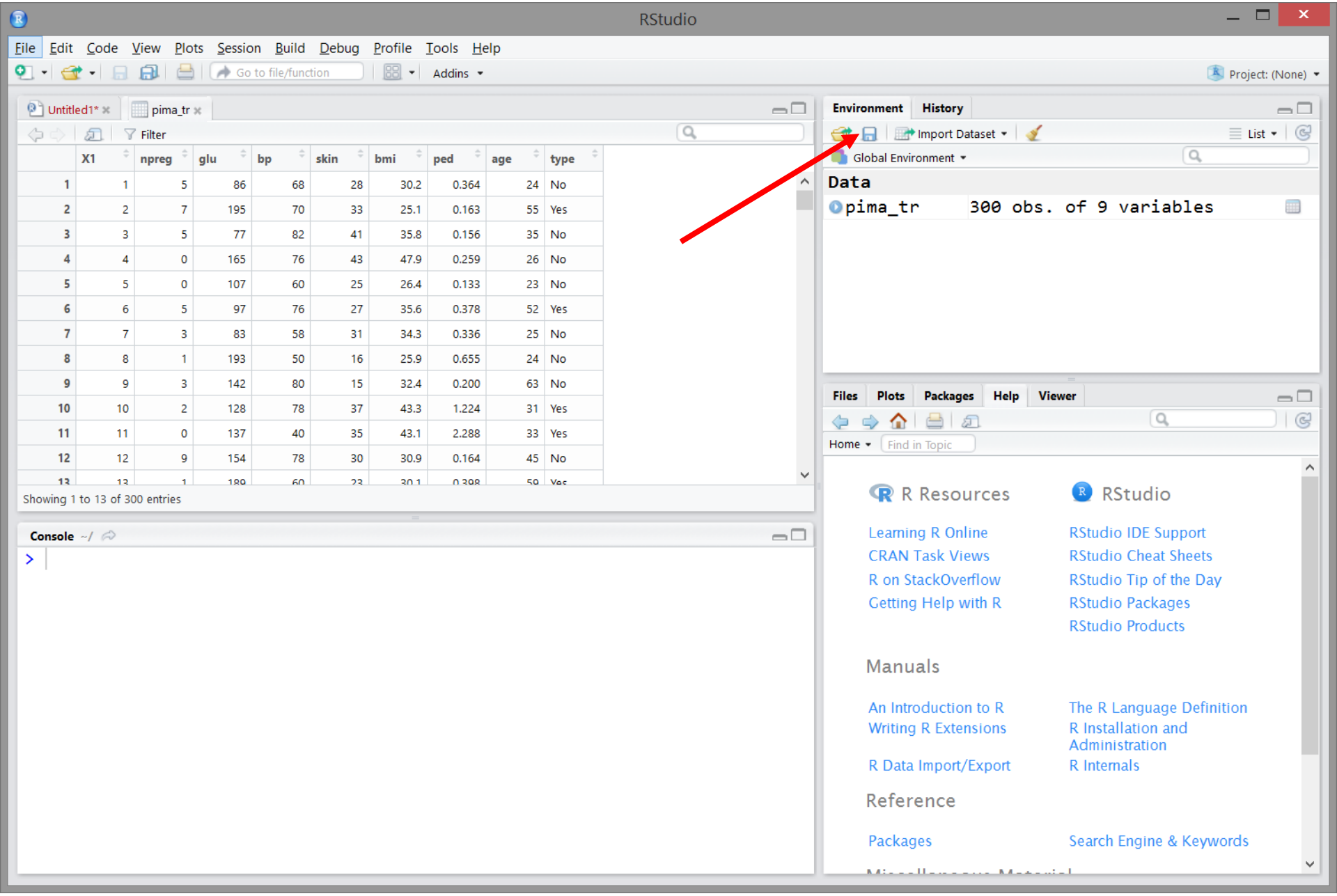
```
setwd("C:\\Users\\Lukas\\Documents\\R intro")

setwd("C:/Users/Lukas/Documents/R intro")
```


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
New Session
Interrupt R
Terminate R...
Restart R Ctrl+Shift+F10
Set Working Directory
Load Workspace...
Save Workspace As...
Clear Workspace...
Quit Session... Ctrl+Q
To Source File Location
To Files Pane Location
Choose Directory... Ctrl+Shift+H
Project: (None)
Untitled1
Source
Run
Source
1
1:1
(Top Level)
R Script
Console ~/
Environment History
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Install
Update
Name
Description
Version
User Library
abind Combine Multidimensional Arrays 1.4-5
acepack ACE and AVAS for Selecting Multiple Regression Transformations 1.4.1
assertthat Easy Pre and Post Assertions 0.2.0
audio Audio Interface for R 0.1-5
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.0
base64enc Tools for base64 encoding 0.1-3
beepr Easily Play Notification Sounds on any Platform 1.2
BH Boost C++ Header Files 1.62.0-1
bindr Parametrized Active Bindings 0.1
bindrcpp An 'Rcpp' Interface to Active Bindings 0.2
bitops Bitwise Operations 1.0-6
Cairo R graphics device using cairo graphics library for creating high-quality bitmap (PNG, JPEG, TIFF), vector (PDF, SVG, PostScript) and display (X11 and Win32) output 1.5-9

Data output

- Save entire workspace
  - Save all R objects you've created so far
  - Allows to return to work/backup current work


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Untitled1*
pima_tr
Filter
X1 npreg glu bp skin bmi ped age type
1 1 5 86 68 28 30.2 0.364 24 No
2 2 7 195 70 33 25.1 0.163 55 Yes
3 3 5 77 82 41 35.8 0.156 35 No
4 4 0 165 76 43 47.9 0.259 26 No
5 5 0 107 60 25 26.4 0.133 23 No
6 6 5 97 76 27 35.6 0.378 52 Yes
7 7 3 83 58 31 34.3 0.336 25 No
8 8 1 193 50 16 25.9 0.655 24 No
9 9 3 142 80 15 32.4 0.200 63 No
10 10 2 128 78 37 43.3 1.224 31 Yes
11 11 0 137 40 35 43.1 2.288 33 Yes
12 12 9 154 78 30 30.9 0.164 45 No
Showing 1 to 13 of 300 entries
Console ~/
Environment History
Import Dataset
List
Global Environment
Data
pima_tr 300 obs. of 9 variables
Files Plots Packages Help Viewer
Home
Find in Topic
R Resources
Learning R Online
CRAN Task Views
R on StackOverflow
Getting Help with R
RStudio
RStudio IDE Support
RStudio Cheat Sheets
RStudio Tip of the Day
RStudio Packages
RStudio Products
Manuals
An Introduction to R
Writing R Extensions
R Data Import/Export
The R Language Definition
R Installation and Administration
R Internals
Reference
Packages
Search Engine & Keywords

Quantitative TA in R

Text analysis in R

- Package “quanteda”
  - Developed by Ken Benoit (LSE)
  - Comprehensive package on text analysis methods
- Package “readtext”
  - Ken Benoit & Adam Obeng
  - Package which allows data import from text sources
  - Easy to work with
- Package “stopwords”
  - Ken Benoit, David Muhr & Kohei Watanabe
  - Package containing various stopwords for different languages

Before we start …

- Open the folder “text_analysis_quanti” folder
- Open script file “text_analysis_1.R” in R Studio

First steps in R

- Install all libraries
  - quanteda
  - readtext
  - stopwords
- Set working directory

```
work.dir <-  "C:\\path\\to\\folder\\"

setwd(work.dir)

library(readtext)
library(quanteda)
library(stopwords)
```

Reading texts into R

- `readtext()` function loads all text files into R
  - Very easy to use – reads everything in the folder
  - Supports various document types
    - TXT
    - PDF
    - DOC
    - Twitter data format JSON
    - …
- Arguments
  - File source
  - Encoding

Reading texts into R

- Encoding
  - Text files are usually stored in certain format
- Consider text "*Príklad zlého kódovania*"
  - ASCII/ISO-8859-1: "*PrÃklad zlÃ©ho kÃ³dovania*"
  - UTF-8: "*Príklad zlého kódovania*"
- As a rule of thumb, UTF-8 encoding is a desired choice

Reading texts into R

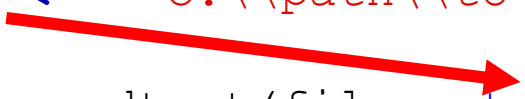
```
text.dir <-  "C:\\path\\to\\folder\\texts\\"

texts <- readtext(file = text.dir, encoding = "UTF-8")
```

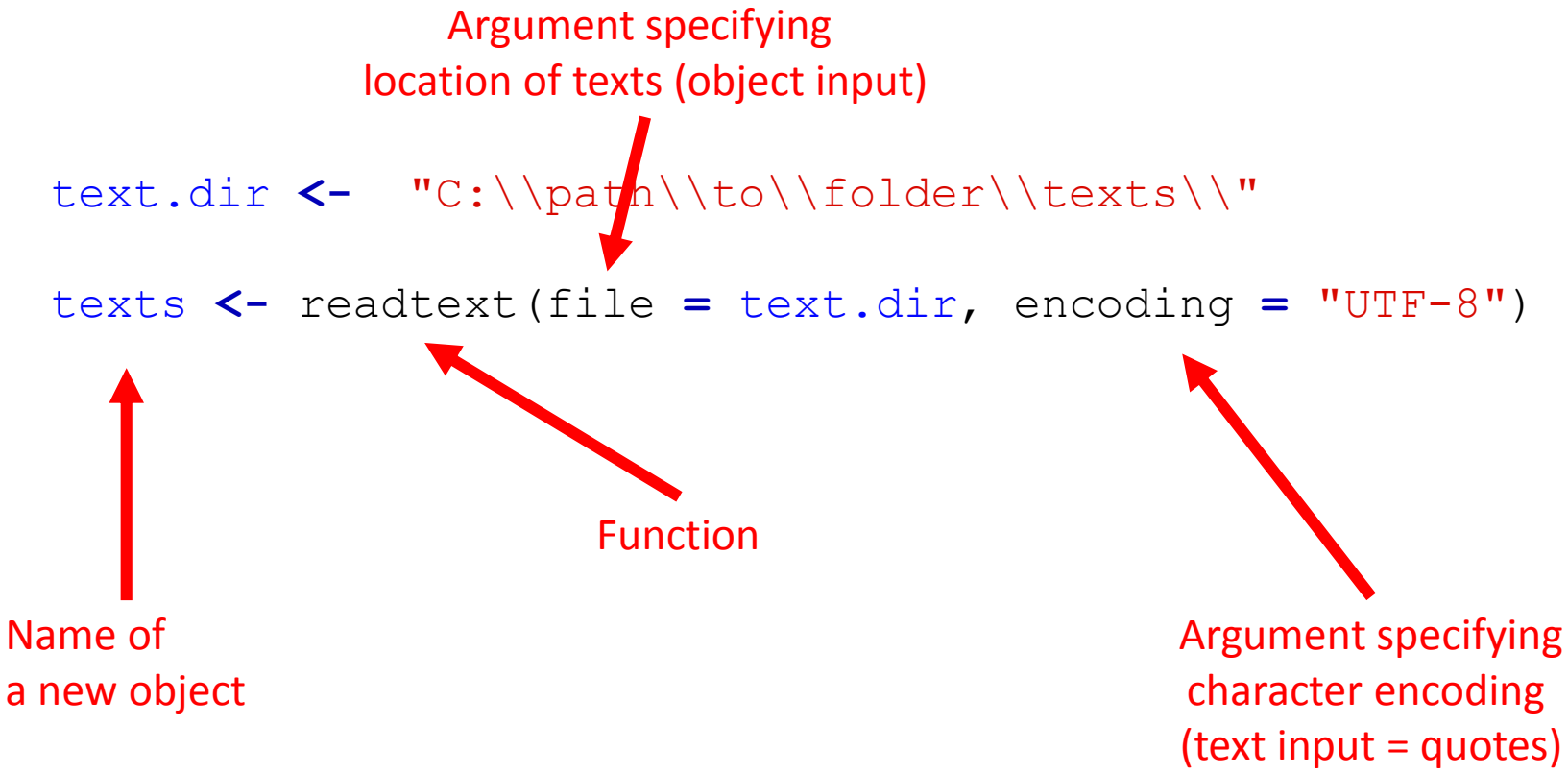
Reading texts into R

```
text.dir <-  "C:\\path\\to\\folder\\texts\\"

texts <- readtext(file = text.dir, encoding = "UTF-8")
```



Reading texts into R



Corpus

- Simple function `corpus()`
  - Creates corpus from all imported texts from the previous step
- All sorts of statistics may be acquired once corpus is generated – e.g. function `summary()`

```
corp <- corpus(x = texts)

ndoc(corp)

summary(corp)
```

Document-feature matrix

- Two-step process in “quanteda” package
- Tokenization of corpus
  - A step necessary to apply some pre-processing choices which are not text-based (removal of noise)
    - Remove numbers
    - Remove punctuation
    - Remove white space (separators)
- Creation of DFM
  - Furthter pre-processing choices
    - Stemming
    - Lowercasing
    - Stopwords removal

Document-feature matrix

- Function `dfm()`
  - Documents in rows
  - Features (tokens) in columns
- Output is in format understood by quanteda

Document-feature matrix

```
tokenization <- tokens(x = corp,
                       remove_numbers = TRUE,
                       remove_punct = TRUE,
                       remove_separators = TRUE,
                       remove_hyphens = FALSE )

doc.term.matrix <- dfm(x = tokenization,
                       tolower = TRUE )
```

Wordcloud

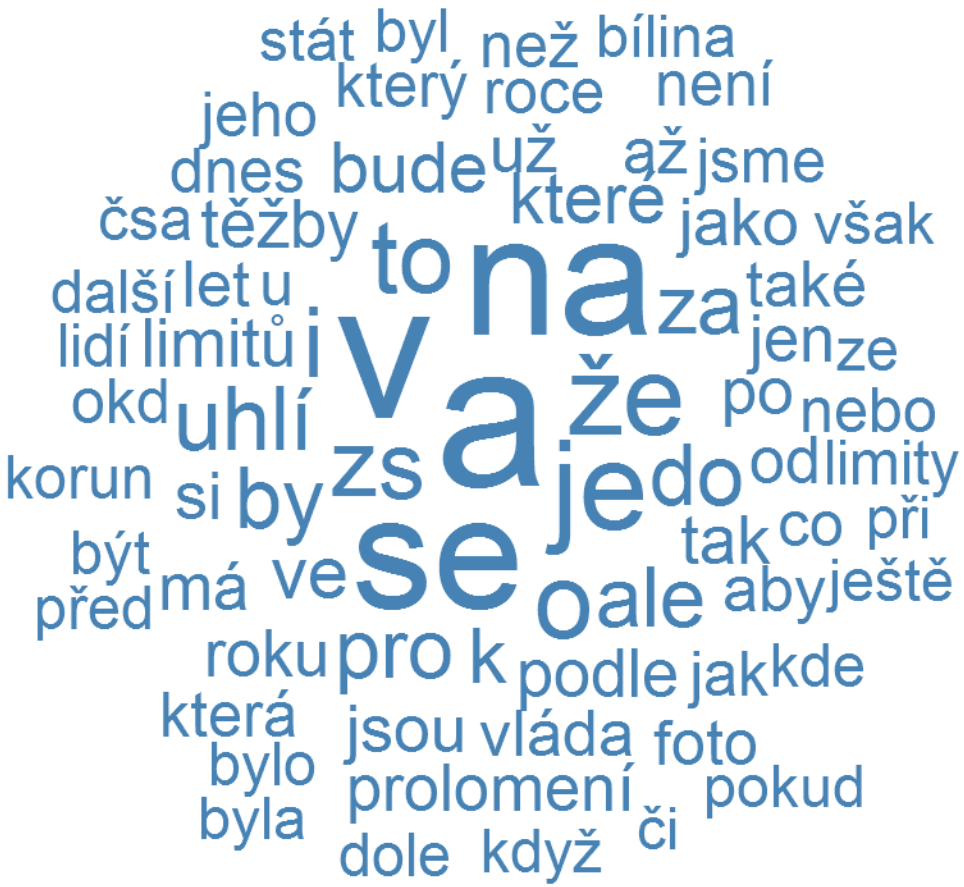
- Function `textplot_wordcloud()`

Attribute	Description
`x`	Terms
`max_words`	Maximum number of words rendered
`min_size`	Size of smallest category
`max_size`	Size of largest category
`rotation`	Percentage of terms placed vertically
`color`	Color or color palette
`...`	Many other arguments available

Wordclouds

```
textplot_wordcloud(x = doc.term.matrix,
                   max_words = 50,
                   min_size = 1,
                   max_size = 4,
                   rotation = 0,
                   color = "steelblue")
```

Wordcloud


stát
byl
než
bílina
který
roce
není
jeho
dnes
bude
už
až
jsme
které
jako
však
čsa
těžby
to
na
za
také
další
let
u
lidí
limitů
i
v
a
že
jen
ze
okd
uhlí
zs
je
do
po
nebo
odlimity
korun
si
by
se
tak
co
při
být
má
ve
o
ale
aby
ještě
před
roku
pro
k
podle
jak
kde
která
jsou
vláda
foto
bylo
prolomení
pokud
byla
či
dole
když