Text analysis 2

Lukáš Lehotský

Useful practicalities

Character encoding

Character encoding

- Character sets
  - Ways how characters in text are translated to code
  - Different standards and character sets
  - Translation from one set to another might result in broken text
- Most common encoding types
  - ASCII – very basic character set – 128 characters
  - Windows 1250 – native Windows CE
  - UTF-8

Character encoding

- ASCII
  - PirÃ¡ti nestojÃ o to, aby prezident MiloÅ¡ Zeman umÄ¡le natahoval vlÃ¡dnutÃ kabinetu bez dÅ¯vÄ¡ry, pokud Å¡Ã©f ANO Andrej BabiÅ¡ neuspÄ¡je se svou menÅ¡inovou vlÃ¡dou doplnÄnou o nestranickÃ© odbornÃky.
- Windows 1250
  - PirĂˇti nestojĂ o to, aby prezident MiloĹˇ Zeman umÄ›le natahoval vlĂˇdnutĂ kabinetu bez dĹŻvÄ›ry, pokud ĹˇĂ©f ANO Andrej BabiĹˇ neuspÄ›je se svou menĹˇinovou vlĂˇdou doplnÄ›nou o nestranickĂ© odbornĂky.
- UTF-8
  - Piráti nestojí o to, aby prezident Miloš Zeman uměle natahoval vládnutí kabinetu bez důvěry, pokud šéf ANO Andrej Babiš neuspěje se svou menšinovou vládou doplněnou o nestranické odborníky.

CSV

CSV

- **C**omma-**S**eparated **V**alues
- Most common table data format
- Data separated by “separator”/“delimiter” (comma, tabulator, space, semicolon,…)
- CSV file (.csv), TSV file (.tsv) – a **text file** (.txt)
- Must have **same number of columns** (separators)

```
cars,type,price,consumption,emissions,expensive
BMW,3,1200000,6.2,0,0
Audi,A4,1164000,5.9,0,0
VW,Passat,950500,6.2,NA,NA
```

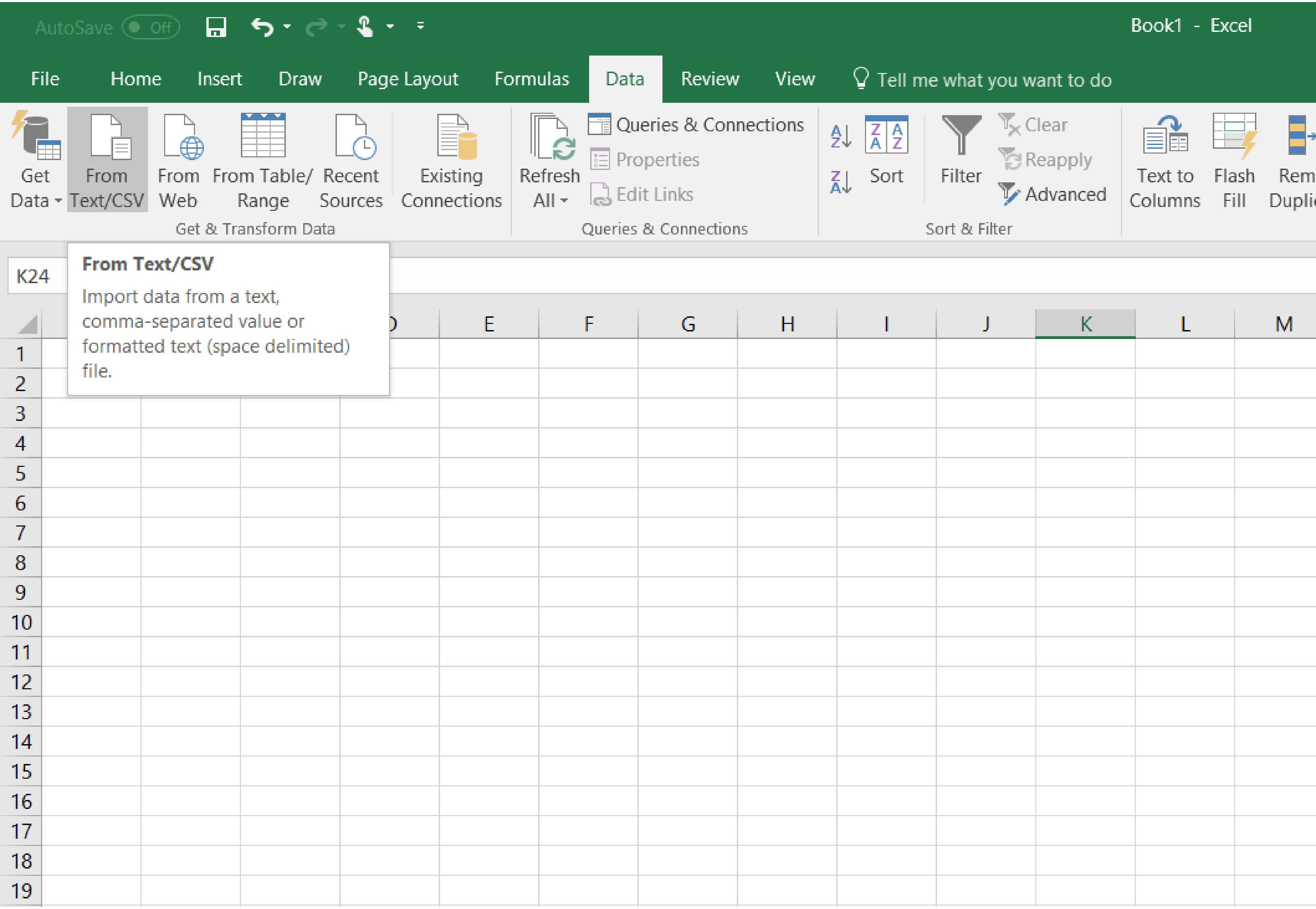
CSV – other examples

```
cars;type;price;consumption;emissions
BMW;3;1200000;6.2;0
Audi;A4;1164000;5.9;0
VW;Passat;950500;6.2;0
```

```
"cars" "type" "price" "consumption"  "emissions"
"BMW" "3" "1,200,000" "6.2" "0"
"Audi" "A4" "1,164,000" "5.9" "0"
"VW" "Passat" "950,500" "6.2" "0"
```

```
cars,type,price,consumption,emissions
BMW,3,1,200,000,6.2
Audi,A4,1,164,000,5.9
VW,Passat,950,500,6.2
```

Bad data – improper use of comma delimiter results in uneven # of rows



From Text/CSV

Import data from a text, comma-separated value or formatted text (space delimited) file.

Topic modeling

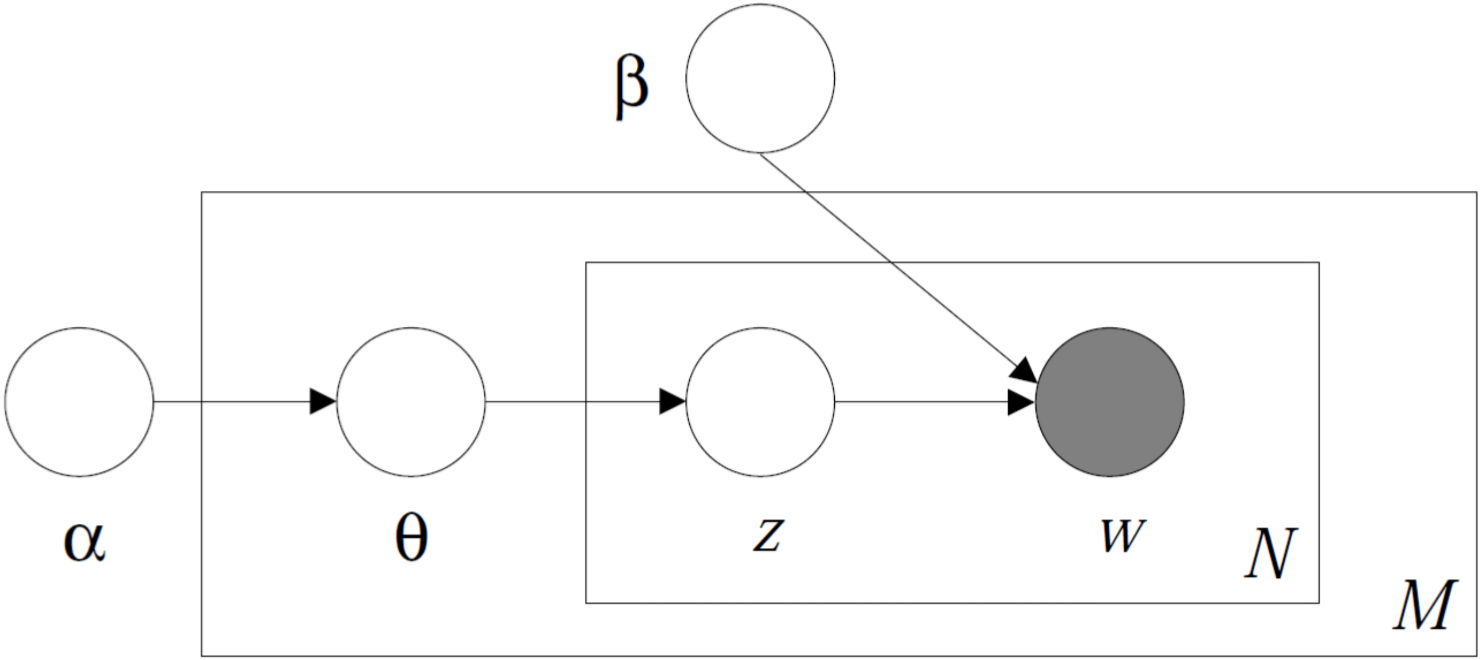
Term-document matrix

	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**

Latent Dirichlet Allocation (LDA)

- Most basic topic model (Blei, Ng & Jordan 2003)
  - Iterative
  - Generative
  - Bayesian
- Documents are an **observed structure**
- **Latent structures (variables)** are underlying documents
  - Structure of topics within documents
  - Structure of words within topics

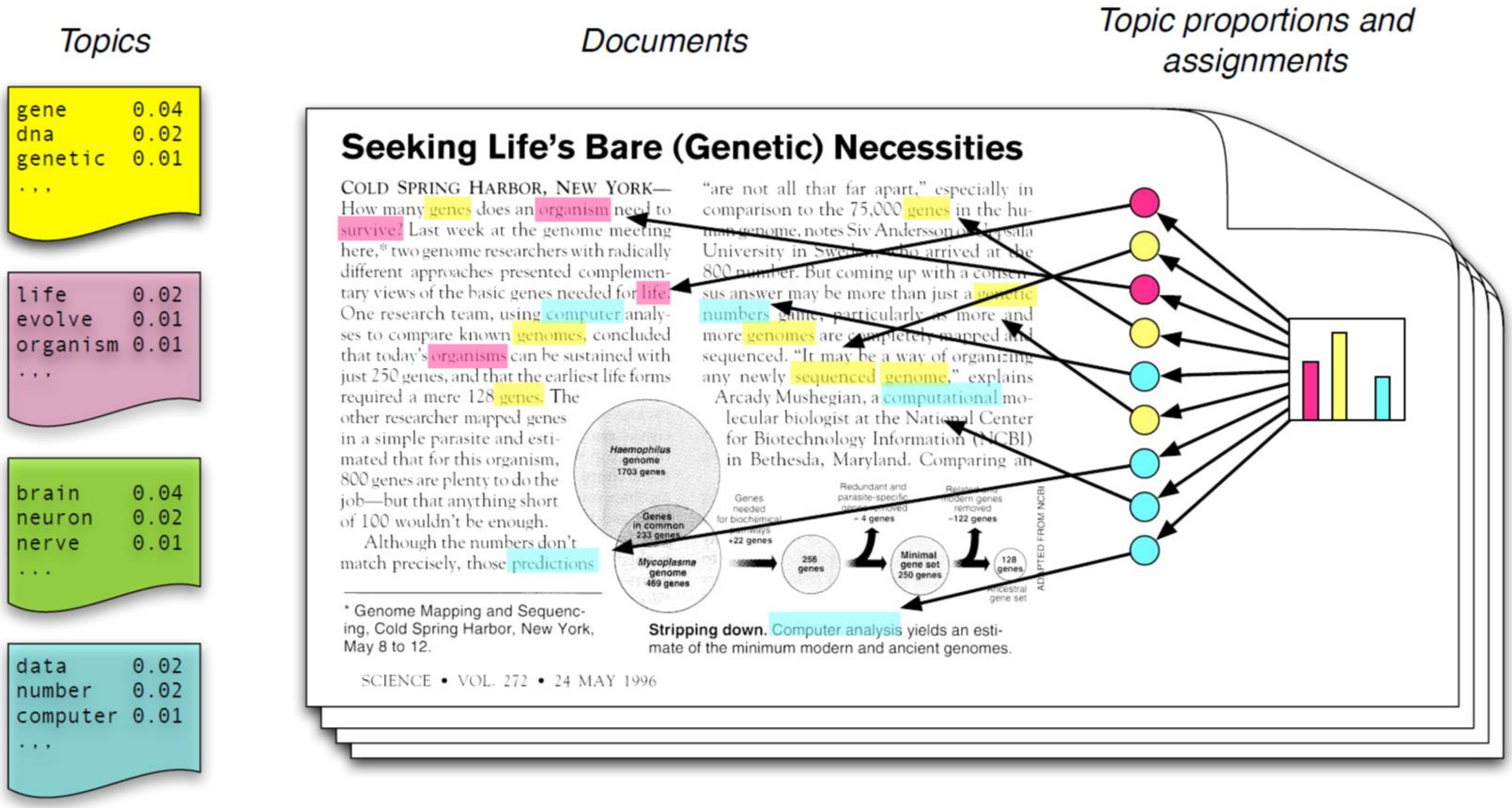
LDA – “plate” scheme



(Blei 2011)

LDA

- Documents are **manifestation of latent variables**
  - Each text is drawn from topics with various probability
  - Each topic is drawn from words with various probability
- Probabilities' base means:
  - **All topics are included** in each document
  - **All words are included** in each topic
  - Presence/absence of topic in document/word in topic expressed by **variation in probability**



gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

(Blei 2011)

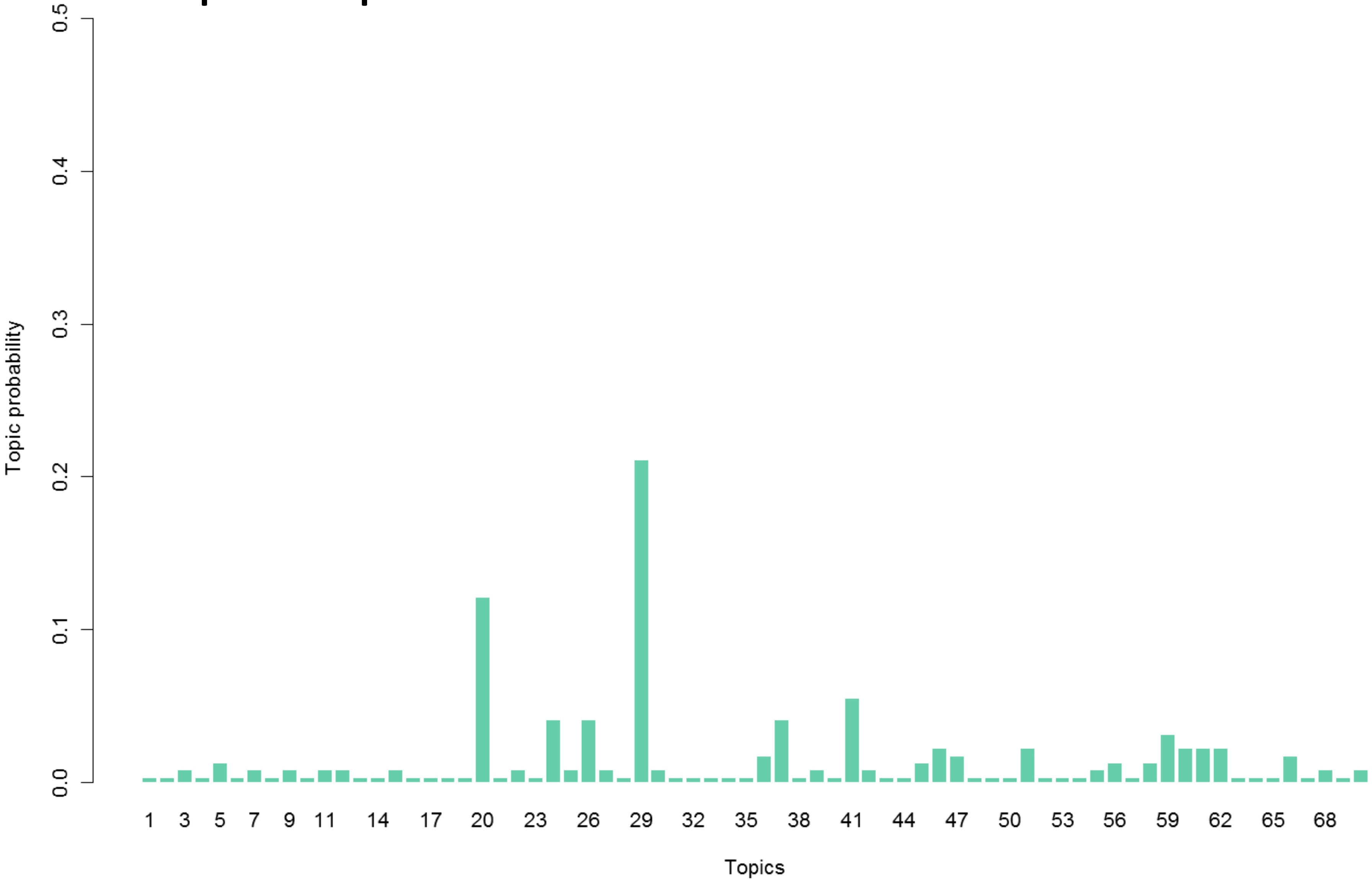
Topic terms – output

Topic.14	Topic.15	Topic.16	Topic.18	Topic.23	Topic.24	Topic.29	Topic.30
firma	firma	horník	muzeum	těžba	zákon	vláda	obec
ředitel	akcie	vláda	výstava	referendum	stát	návrh	obyvatel
velký	skupina	útlum	návštěvník	litvínov	horní	ministr	horní_jiřetín
podnik	podíl	odborář	divadlo	uhlí	novela	ministerstvo	těžba
společnost	investice	odbory	akce	aktivista	pozemek	minpo	dům
zaměstnanec	obchod	sociální	expozice	greenpeace	návrh	koncepce	černice
patřit	j_and_t	těžba	dítě	zastupitel	změna	mpo	starosta
závod	investor	odborový	otevřít	akce	možnost	počítat	jiřetín
zakázka	euro	předseda	film	sdružení	muset	sek	vesnice
představenstvo	eph	důl	konat	zachování	pspčr	varianta	mus
vedení	prodej	stávka	galerie	ekolog	majitel	průmysl	horní

Topics per document distribution

Topic	Document 250-1997-06-11
29	0.210916
20	0.121294
41	0.055256
24	0.041105
26	0.041105
37	0.041105
59	0.031671
46	0.022237
51	0.022237
60	0.022237
…	…

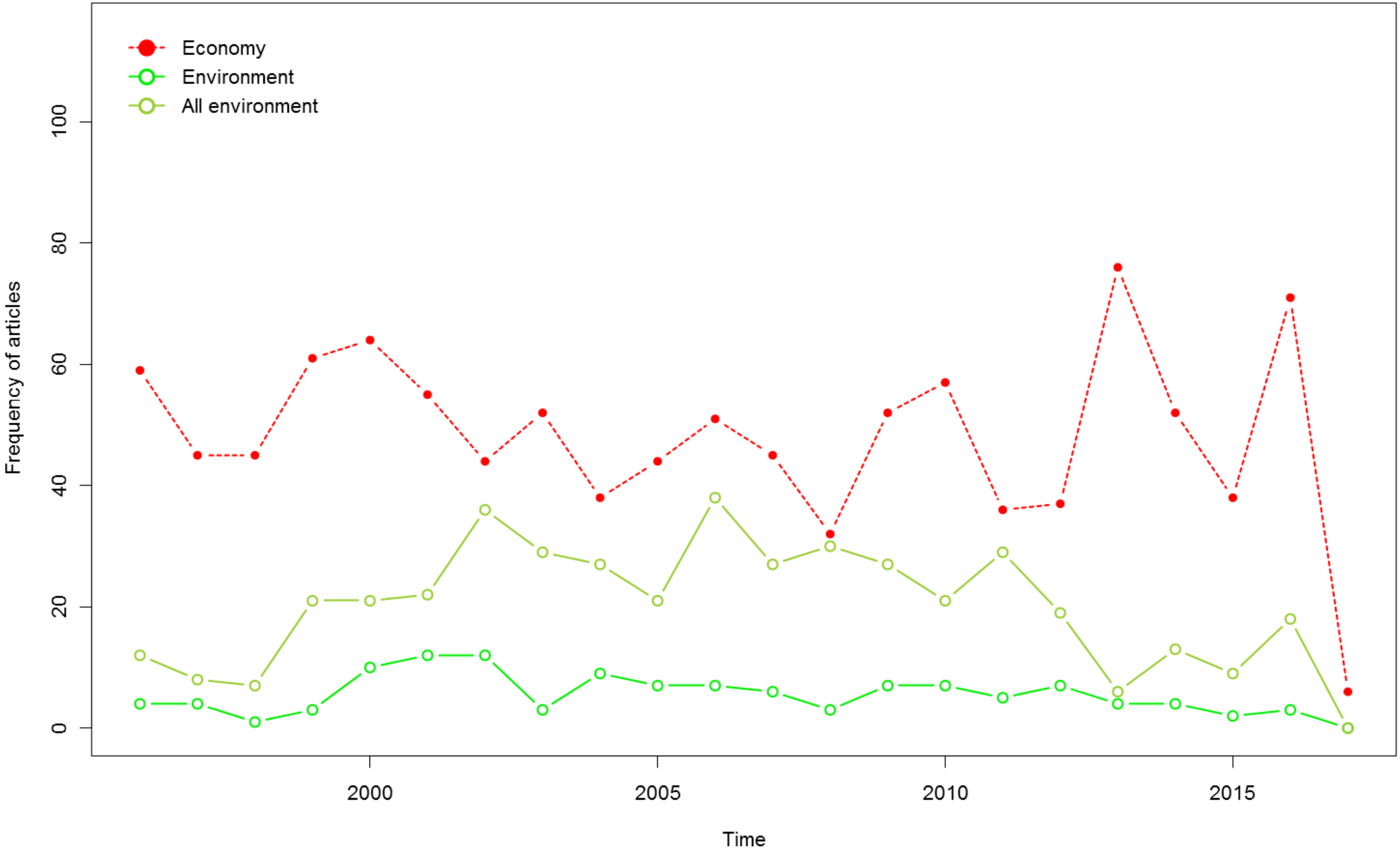
Topics per document distribution



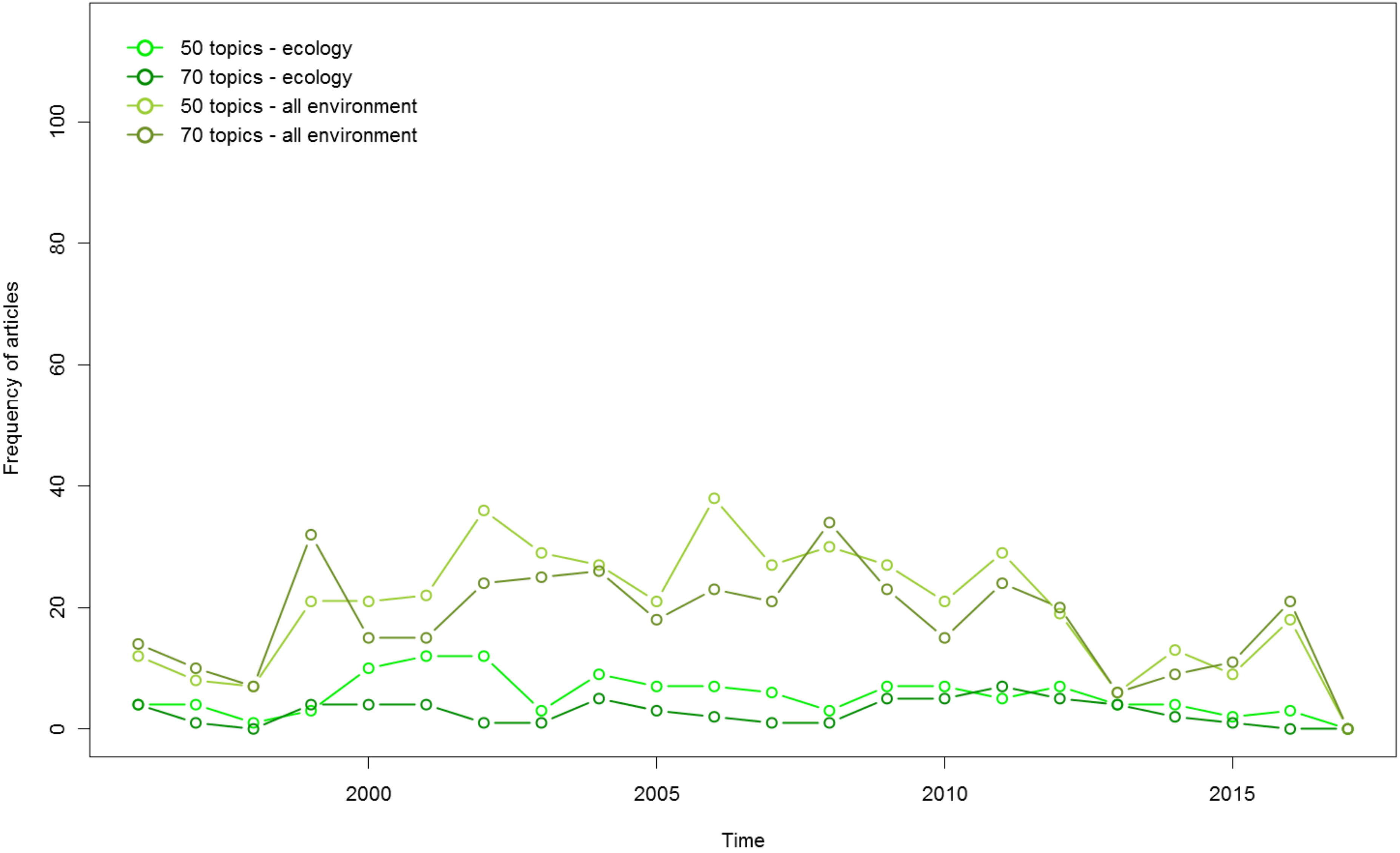
Topics per document distribution

	250-1997-06-11	**251-1997-06-12**	**252-1997-06-13**	**253-1997-06-13**	**254-1997-06-14**	**255-1997-06-17**
4	0.003369	0.006676	0.008682	0.012864	0.027405	0.008487
5	0.012803	0.006676	0.054722	0.017603	0.002915	0.003536
6	0.003369	0.053405	0.001315	0.003385	0.01516	0.008487
7	0.008086	0.006676	0.019732	0.008125	0.002915	0.072843
8	0.003369	0.016021	0.014207	0.008125	0.002915	0.152051
9	0.008086	0.006676	0.001315	0.008125	0.019242	0.048091
10	0.003369	0.006676	0.010524	0.003385	0.006997	0.008487

Topics per document distribution



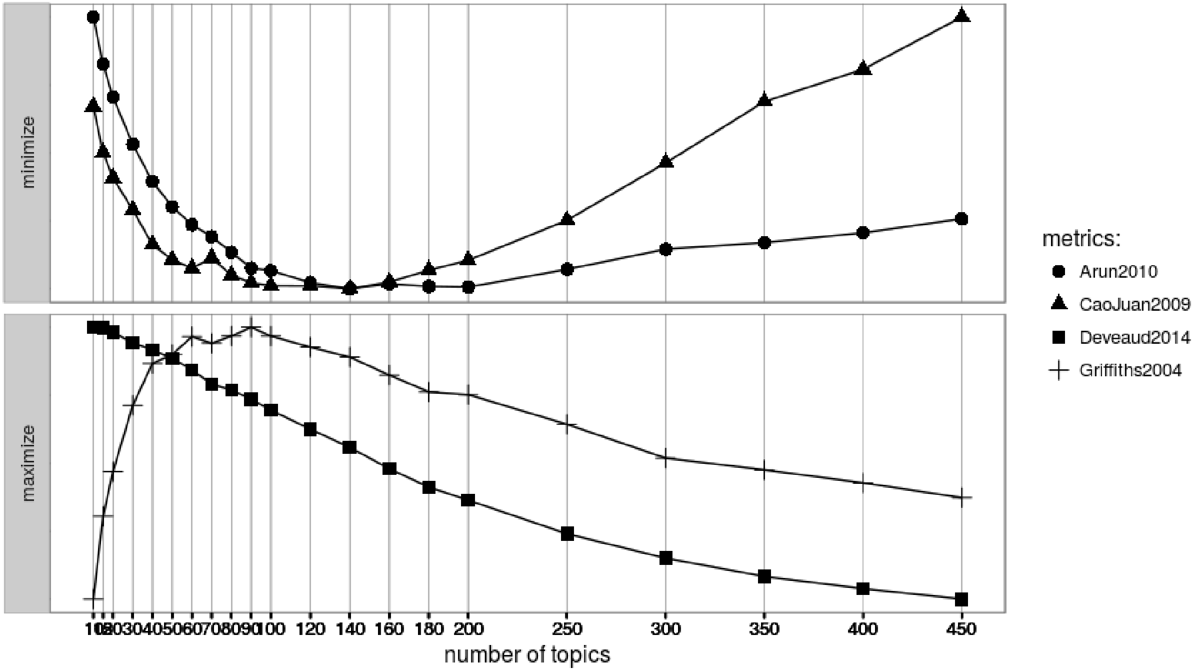
Topics per document distribution



Limits

- **Large number of documents** necessary to approximate a good model
- If many topics in documents, LDA **approximates worse**
- **Not suitable** for analysis of **short texts** (Twitter, Facebook)
- Based on **bag-of-words** assumption
- **Number of topics** has to be established **prior** to the analysis
- Fit vs. **interpretability** (Chang 2009)

Selecting number of topics

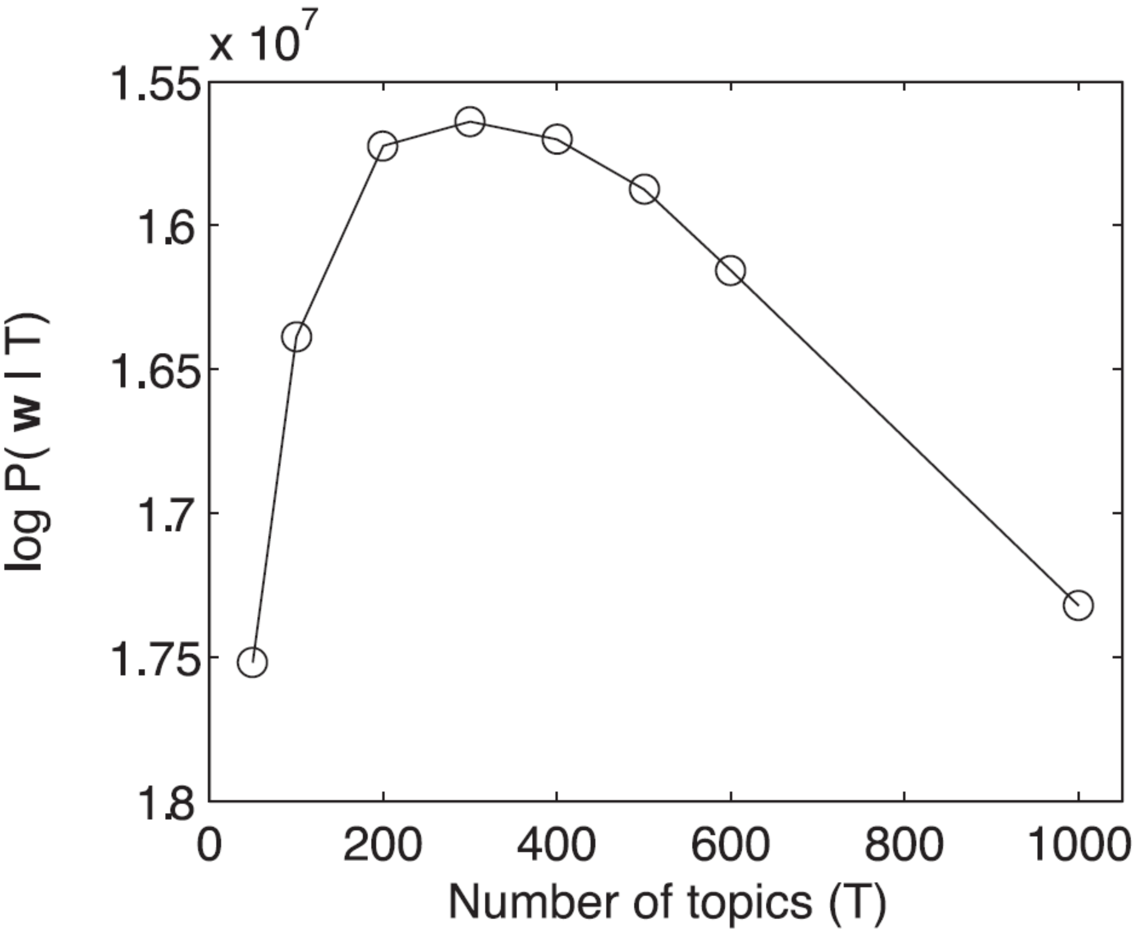


(Murzintcev 2016)

Selecting number of topics

- Mathematical fit
  - Fit of the model comparable to **fit of any other quantitative model**
  - Addressing how well the model describes existing data
  - **Perplexity** (Griffiths & Styvers 2004)
    - Corpus split into 2 parts
    - Model approximates on one set (training set)
    - Perplexity measured on the other set (test set)
    - Lower perplexity = better model fit
  - Overfitting

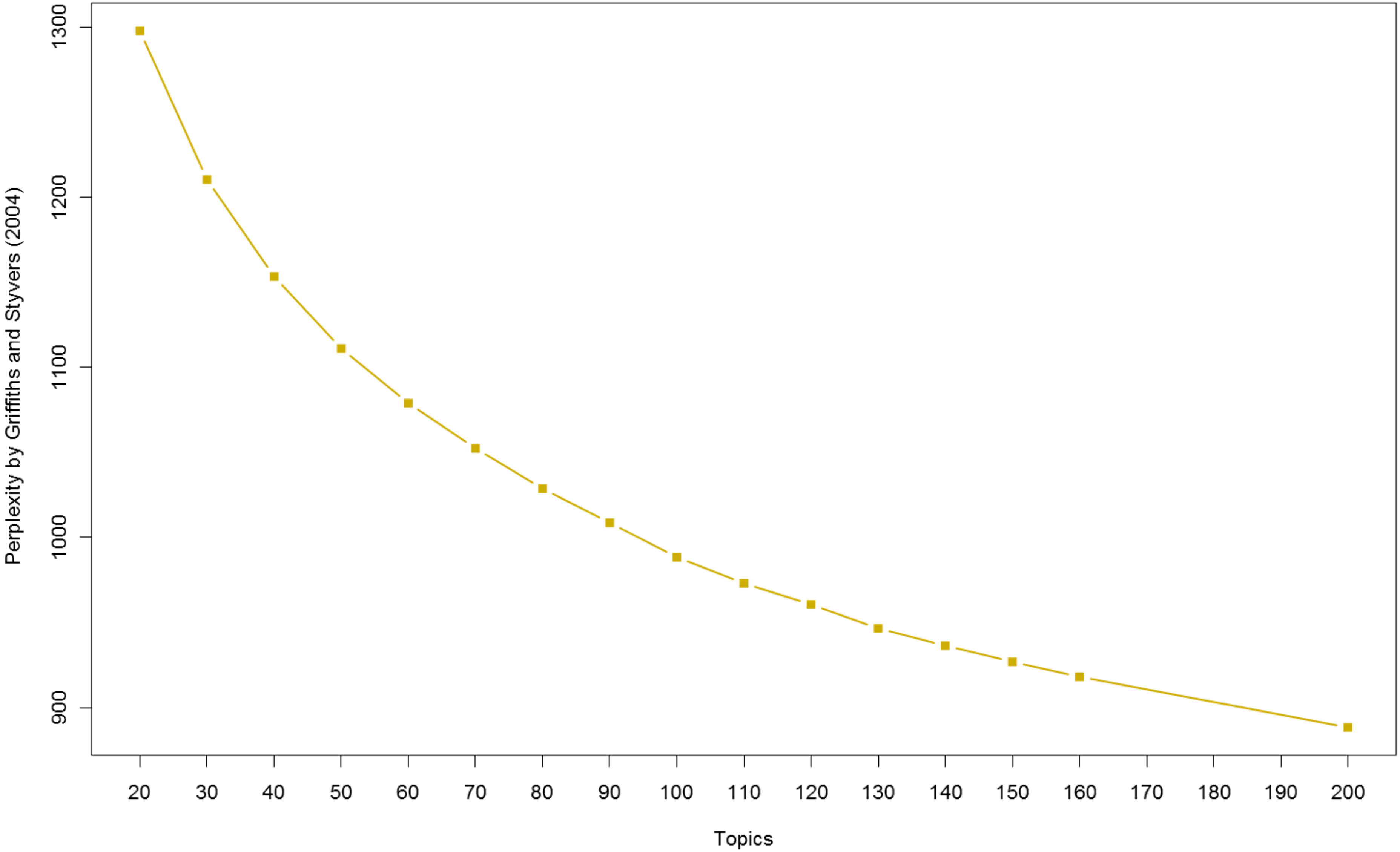
Selecting number of topcis



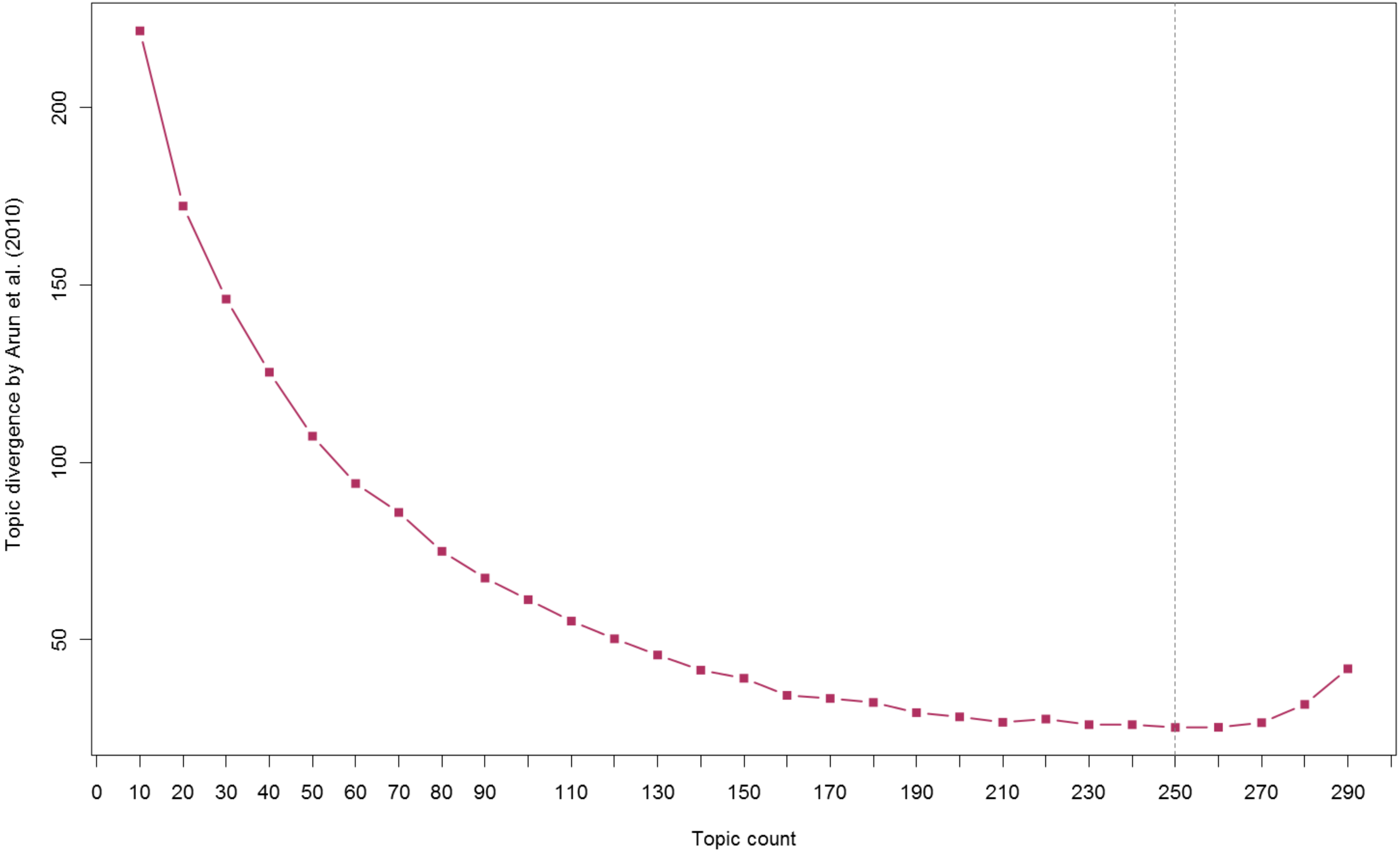
Selecting number of topics

- Other measures
  - Based on divergence of topics
- Arun et al. 2010
  - Matrix factorization and Kullback-Leibler divergence (relative entropy)
- Cao et al. 2009
  - Cosine similarity between topics – based on terms and their positions

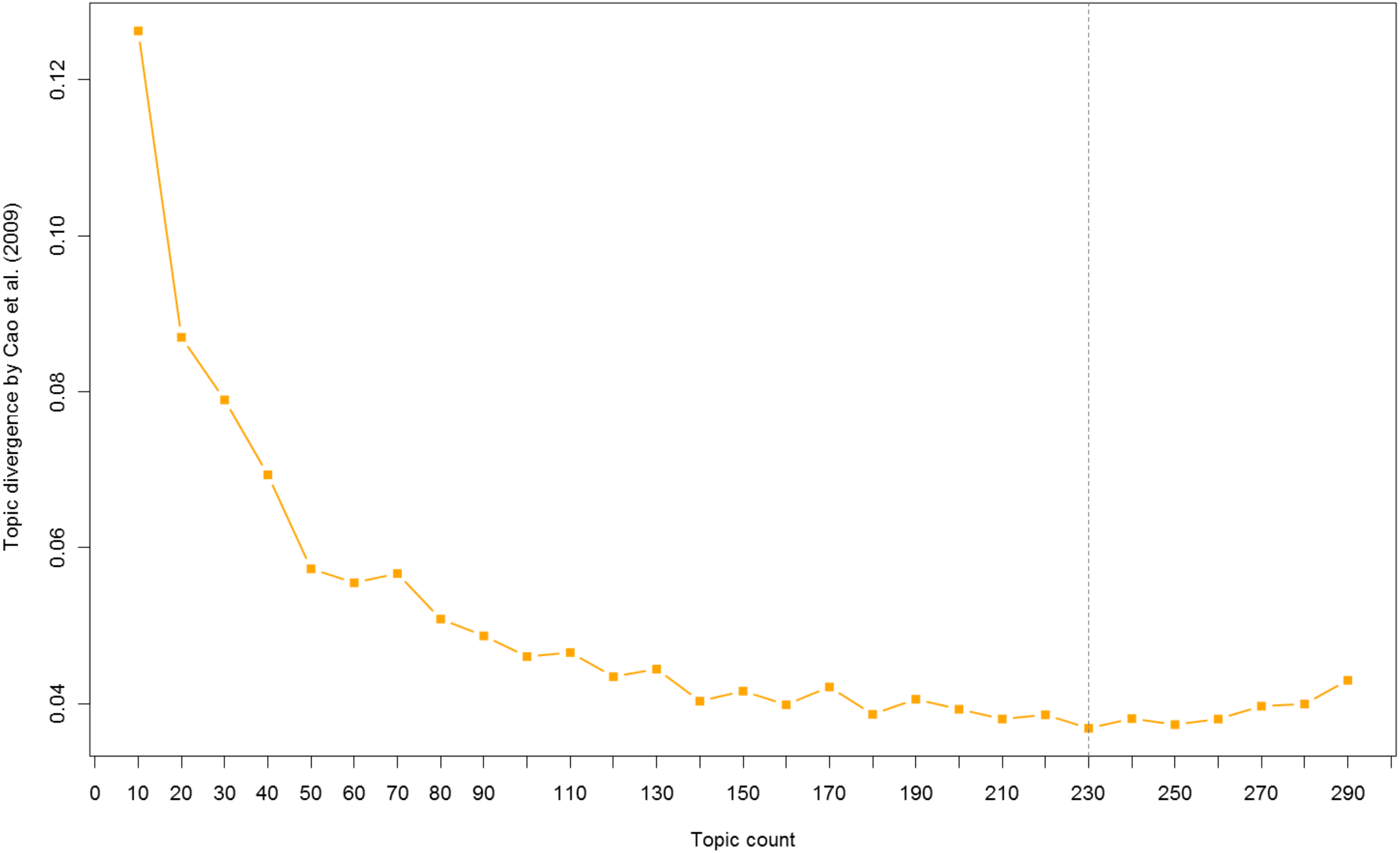
Perplexity


Perplexity by Griffiths and Styvers (2004)
900
1000
1100
1200
1300
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
Topics

Divergence (Arun et al. 2010)



Divergence (Cao et al. 2009)



Interpretability

- Chang et al. 2009
- Model fit **does not ensure better interpretability** of topics
- Experimental design
    - Most probable terms obtained
    - Among these terms, one random term is inserted in random position (“intruder term”)
    - Human coders asked to identify intruder terms

Interpretability

Topic.8	Topic.9	Topic.10	Topic.11	Topic.12	Topic.13	Topic.14	Topic.15
limit	názor	stavba	město	politika	elektrárna	důl	zámek
těžba	otázka	silnice	ostrava	ekonomika	jaderný	uhlí	památka
uhlí	problém	doprava	karviná	ekonomický	elektřina	strop	kostel
prolomení	občan	výnosný	starý	úroveň	temelín	horník	hrad
horní_jiřetín	zájem	praha	centrum	stát	energetika	důlní	muzeum
černice	stát	český	zóna	země	darkov	hornický	brzký
minulost	životní_prostředí	dálnice	průmyslový	oblast	energetický	uhelný	starý
litvínov	informace	vést	generální	objednat	čez	těžit	svatý
obec	muset	dráha	radnice	vývoj	blok	kladno	stát
jiřetín	případ	železniční	ostravský	státní	výstavba	hornictví	objekt
hnědý_uhlí	jít	cesta	městský	zahraniční	zdroj	těžba	areál

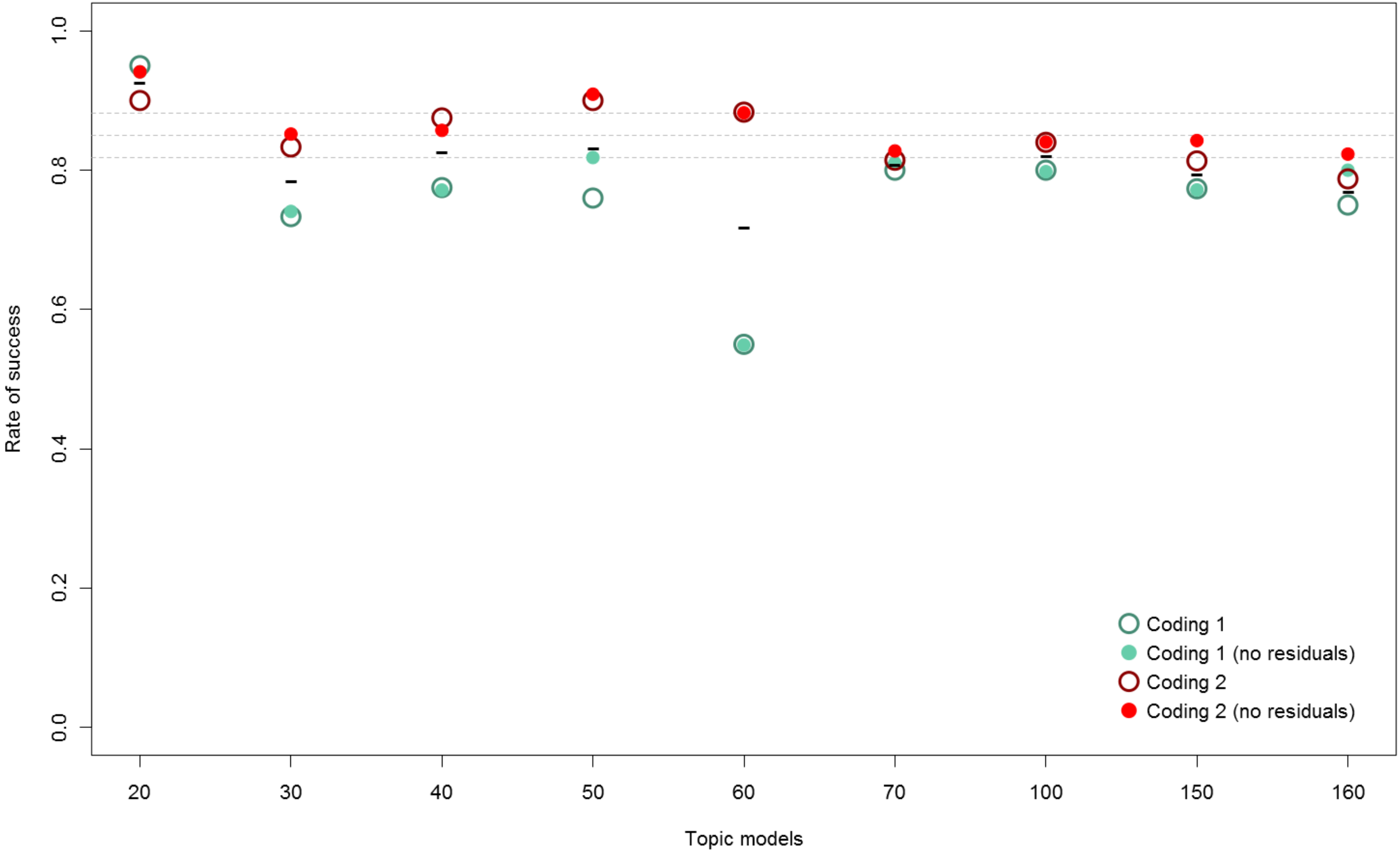
Interpretability

Topic.8	Topic.9	Topic.10	Topic.11	Topic.12	Topic.13	Topic.14	Topic.15
limit	názor	stavba	město	politika	elektrárna	důl	zámek
těžba	otázka	silnice	ostrava	ekonomika	jaderný	uhlí	památka
uhlí	problém	doprava	karviná	ekonomický	elektřina	strop	kostel
prolomení	občan	výnosný	starý	úroveň	temelín	horník	hrad
horní_jiřetín	zájem	praha	centrum	stát	energetika	důlní	muzeum
černice	stát	český	zóna	země	darkov	hornický	brzký
minulost	životní_prostředí	dálnice	průmyslový	oblast	energetický	uhelný	starý
litvínov	informace	vést	generální	objednat	čez	těžit	svatý
obec	muset	dráha	radnice	vývoj	blok	kladno	stát
jiřetín	případ	železniční	ostravský	státní	výstavba	hornictví	objekt
hnědý_uhlí	jít	cesta	městský	zahraniční	zdroj	těžba	areál

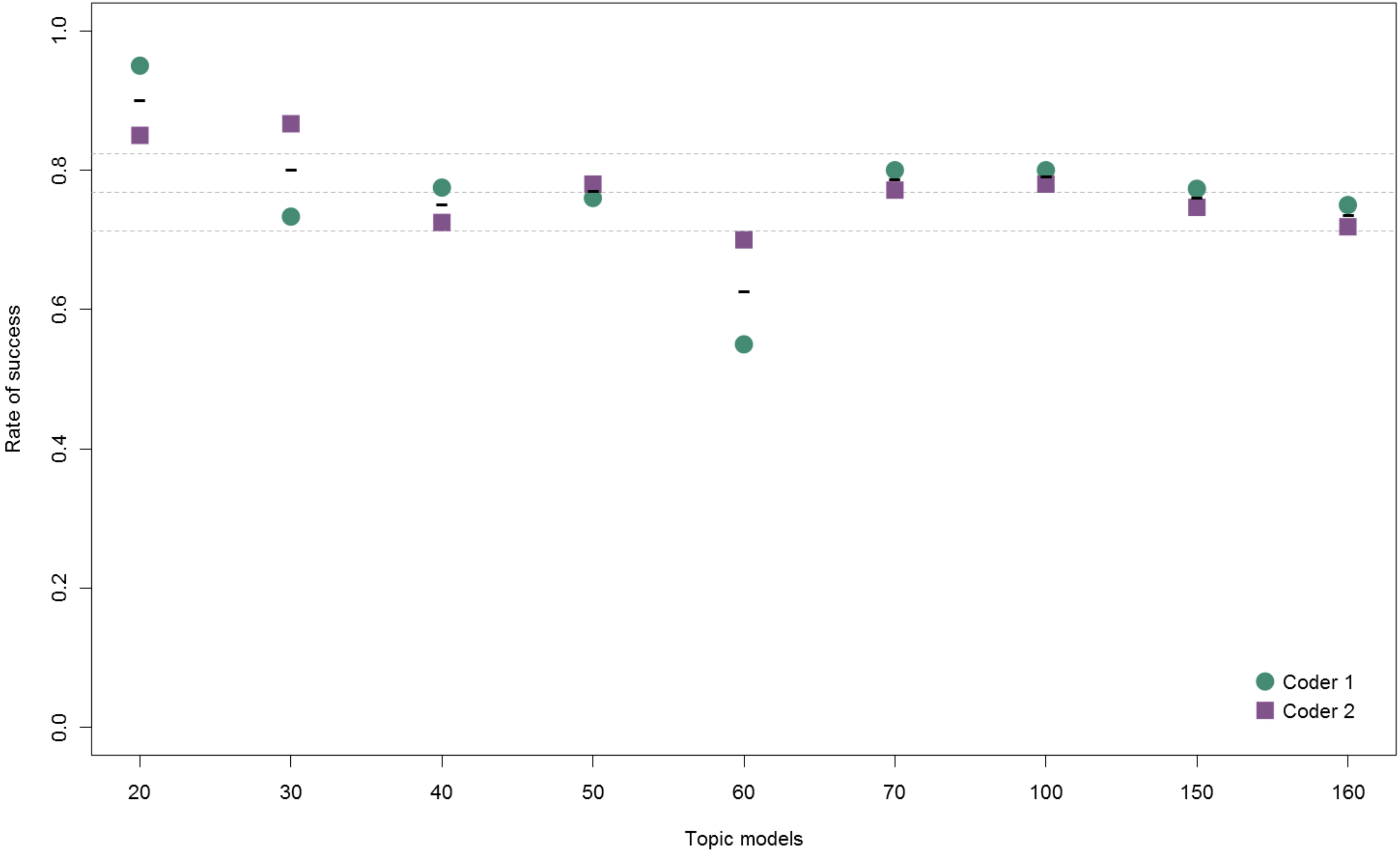
Interpretability

Topic.8	Topic.9	Topic.10	Topic.11	Topic.12	Topic.13	Topic.14	Topic.15
limit	názor	stavba	město	politika	elektrárna	důl	zámek
těžba	otázka	silnice	ostrava	ekonomika	jaderný	uhlí	památka
uhlí	problém	doprava	karviná	ekonomický	elektřina	[redacted]	kostel
prolomení	občan	[redacted]	starý	úroveň	temelín	horník	hrad
horní_jiřetín	zájem	praha	centrum	stát	energetika	důlní	muzeum
černice	stát	český	zóna	země	[redacted]	hornický	[redacted]
[redacted]	životní_prostředí	dálnice	průmyslový	oblast	energetický	uhelný	starý
litvínov	[redacted]	vést	[redacted]	[redacted]	čez	těžit	svatý
obec	muset	dráha	radnice	vývoj	blok	kladno	stát
jiřetín	případ	železniční	ostravský	státní	výstavba	hornictví	objekt
hnědý_uhlí	jít	cesta	městský	zahraniční	zdroj	těžba	areál

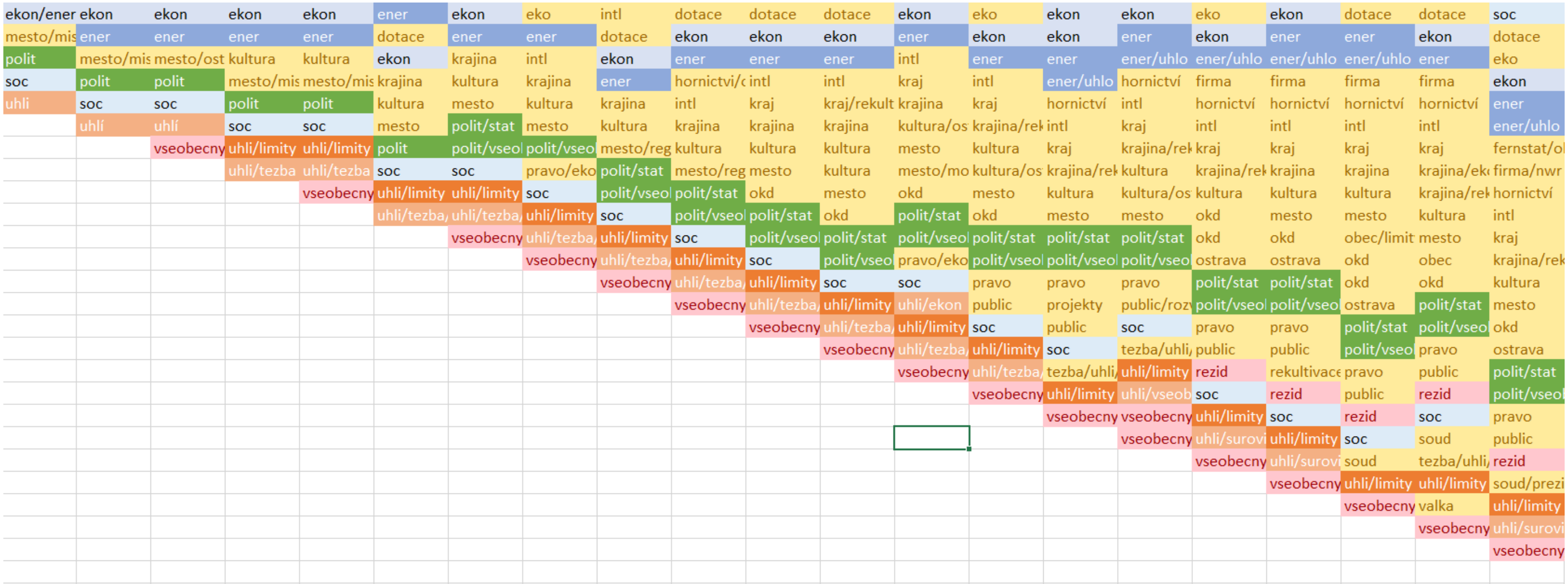
Intercoder performance



Intracoder performance



Topic consistence



ekon/ener	ekon	ekon	ekon	ekon	ener	ekon	eko	intl	dotace	dotace	dotace	ekon	eko	ekon	ekon	eko	ekon	dotace	dotace	soc
mesto/mis	ener	ener	ener	ener	dotace	ener	ener	dotace	ekon	ekon	ekon	ener	ekon	ekon	ener	ekon	ener	ener	ekon	dotace
polit	mesto/mis	mesto/ost	kultura	kultura	ekon	krajina	intl	ekon	ener	ener	ener	intl	ener	ener	ener/uhlo	ener/uhlo	ener/uhlo	ener/uhlo	ener	eko
soc	polit	polit	mesto/mis	mesto/mis	krajina	kultura	krajina	ener	hornictvi/c	intl	intl	kraj	intl	ener/uhlo	hornictví	firma	firma	firma	firma	ekon
uhli	soc	soc	polit	polit	kultura	mesto	kultura	krajina	intl	kraj	kraj/rekult	krajina	kraj	hornictví	intl	hornictví	hornictví	hornictví	hornictví	ener
	uhlí	uhlí	soc	soc	mesto	polit/stat	mesto	kultura	krajina	krajina	krajina	kultura/os	krajina/rek	intl	kraj	intl	intl	intl	intl	ener/uhlo
		vseobecny	uhli/limity	uhli/limity	polit	polit/vseo	polit/vseo	mesto/reg	kultura	kultura	kultura	mesto	kultura	kraj	krajina/rek	kraj	kraj	kraj	kraj	fernstat/o
			uhli/tezba	uhli/tezba	soc	soc	pravo/eko	polit/stat	mesto/reg	mesto	kultura	mesto/mo	kultura/os	krajina/rek	kultura	krajina/rek	krajina	krajina	krajina/ek	firma/nwr
				vseobecny	uhli/limity	uhli/limity	soc	polit/vseo	polit/stat	okd	mesto	okd	mesto	kultura	kultura/os	kultura	kultura	kultura	krajina/rek	hornictví
					uhli/tezba/	uhli/tezba/	uhli/limity	soc	polit/vseo	polit/stat	okd	polit/stat	okd	mesto	mesto	okd	mesto	mesto	kultura	intl
						vseobecny	uhli/tezba/	uhli/limity	soc	polit/vseo	polit/stat	polit/vseo	polit/stat	polit/stat	polit/stat	okd	okd	obec/limit	mesto	kraj
							vseobecny	uhli/tezba/	uhli/limity	soc	polit/vseo	pravo/eko	polit/vseo	polit/vseo	polit/vseo	ostrava	ostrava	okd	obec	krajina/rek
								vseobecny	uhli/tezba/	uhli/limity	soc	soc	pravo	pravo	pravo	polit/stat	polit/stat	okd	okd	kultura
									vseobecny	uhli/tezba/	uhli/limity	uhli/ekon	public	projekty	public/roz	polit/vseo	polit/vseo	ostrava	polit/stat	mesto
										vseobecny	uhli/tezba/	uhli/limity	soc	public	soc	pravo	pravo	polit/stat	polit/vseo	okd
											vseobecny	uhli/tezba/	uhli/limity	soc	tezba/uhli/	public	public	polit/vseo	pravo	ostrava
												vseobecny	uhli/tezba/	tezba/uhli/	uhli/limity	rezid	rekultivace	pravo	public	polit/stat
													vseobecny	uhli/limity	uhli/vseob	soc	rezid	public	rezid	polit/vseo
														vseobecny	vseobecny	uhli/limity	soc	rezid	soc	pravo
															vseobecny	uhli/surovi	uhli/limity	soc	soud	public
																vseobecny	uhli/surovi	soud	tezba/uhli/	rezid
																	vseobecny	uhli/limity	uhli/limity	soud/prezi
																		vseobecny	valka	uhli/limity
																			vseobecny	uhli/surovi
																				vseobecny

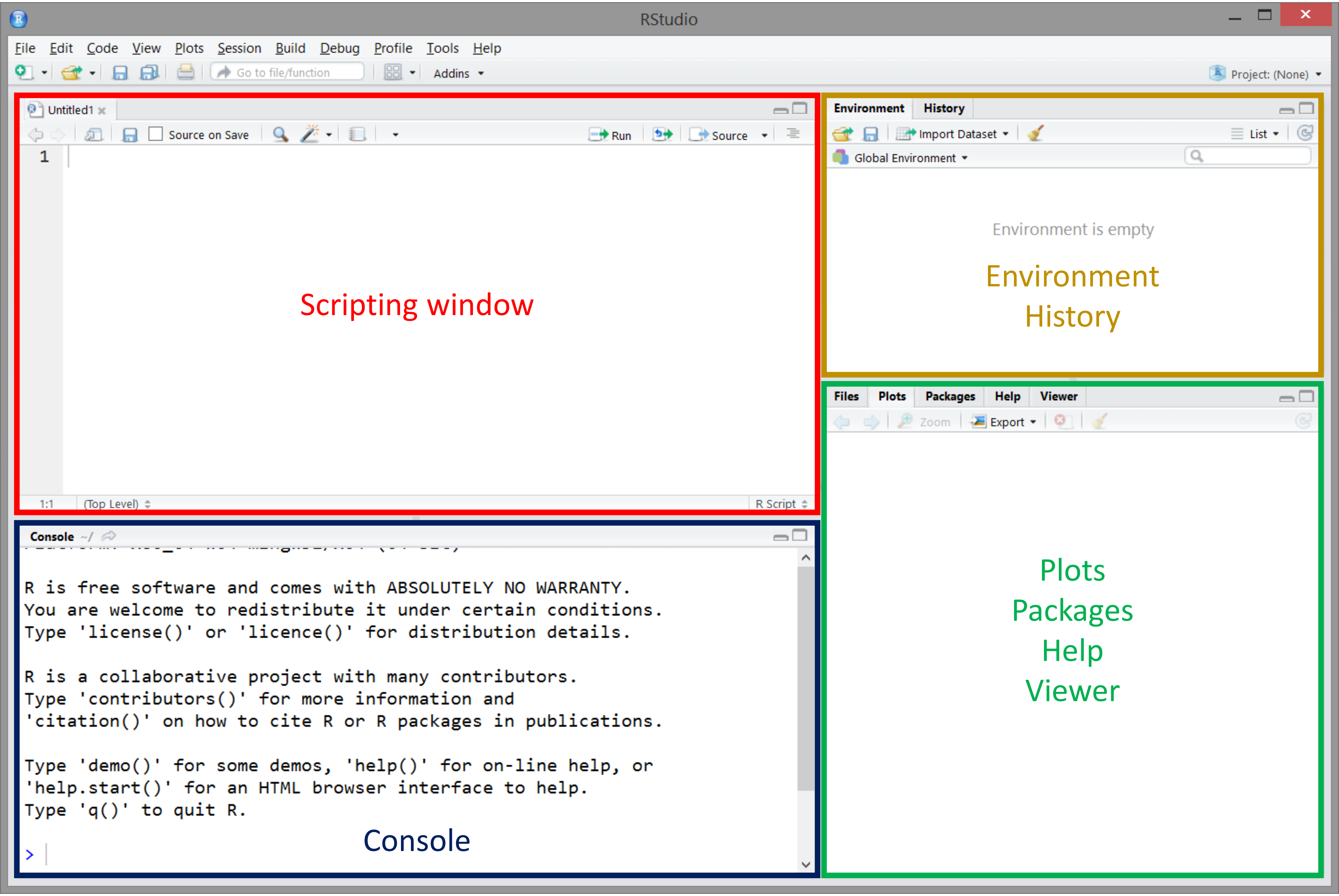
Bad vs. good topics

- Chuang et al. 2013
- Measure **optimization of parameters**
  - Try to avoid **bad topics**
  - Compare performance of topic models against qualitatively constructed concepts
  - Classify **bad topics** as
    - Junk
    - Fused
    - Missing
    - Repeated

Topic model set-up

- Parameters
  - Alpha
    - Determines how “consistent” or “focused” topics are
    - Lower alpha → more “pointed” topics
    - Optimization of parameter (Chuang et al. 2013)
  - Sampling parameters (Gibbs)
    - Number of iterations (how many iterations model does)
    - Omitted iterations from the start
    - Omitted iterations between samples

Programming in R



```
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.

>
```

Help!

- Functions have extensive help descriptions
  - Attributes
  - Examples
  - Citations (very useful)
  - …
- Function `help()`
- Prefix `?` before name of any function

```
help(c)

?c
```

Object types – prop. of objects

- Vector
  - Sequence (1-dimensional) of elements of same data class
- **Matrix**
  - 2-dimensional rectangular collection of elements of same data class
  - Array: n-dimensional matrix
- **List**
  - Vector that can contain elements of different data classes
- Data frame
  - List of vectors of equal length
  - Table data

Vector

```
c(2,3,5)

[1] 2 3 5

c("aa", "bb", "cc", "dd", "ee")

[1] "aa" "bb" "cc" "dd" "ee“

c(TRUE, FALSE, TRUE, FALSE, FALSE)

[1] TRUE FALSE TRUE FALSE FALSE
```

Matrix

```
m <- matrix(data = c(1,2,3,4,5,6,7,8,9,10,11,12),
            nrow = 3,
            ncol = 4)

m
     [,1]  [,2]  [,3]  [,4]
[1,] 1     4     7     10
[2,] 2     5     8     11
[3,] 3     6     9     12
```

List

```
numbers <- c(2, 3, 5)
strings <- c("aa", "bb", "cc", "dd", "ee")
my.list <- list(numbers, strings, 3)

my.list

[[1]]
[1] 2 3 5

[[2]]
[1] "aa" "bb" "cc" "dd" "ee"

[[3]]
[1] 3
```

Data frame

```r
teams <- c("PHI","NYM","FLA","ATL","WSN")
wins <- c(92,89,94,72,59)
losses <- c(70,73,77,90,102)

table.data <- data.frame(teams, wins, losses)

table.data
```

```
      teams wins  losses
1     PHI   92    70
2     NYM   89    73
3     FLA   94    77
4     ATL   72    90
5     WSN   59    102
```

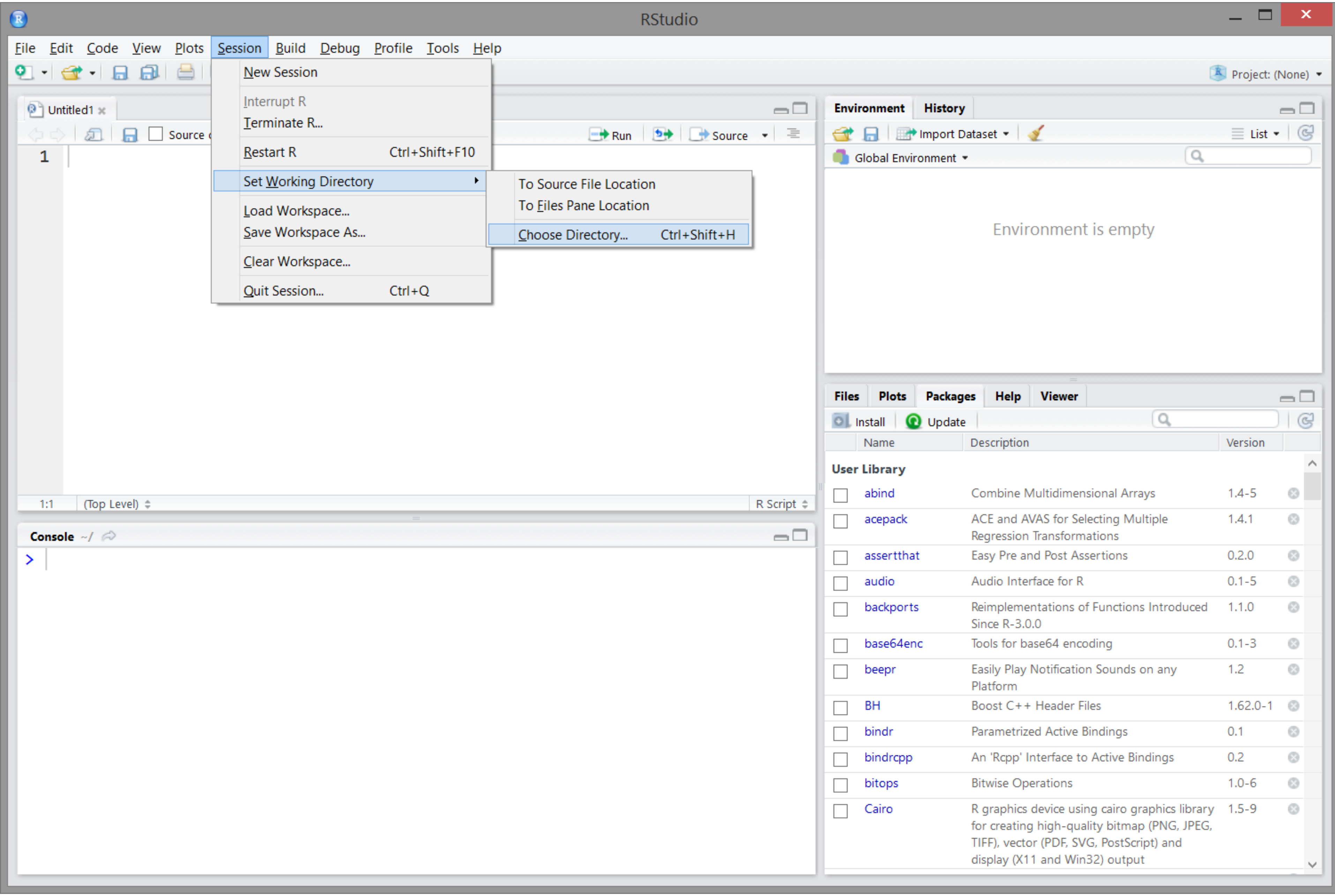
Basic R functions

```
c() # combine two or more elements into an object

class() # explore elements’ data class
length() # explore number of first dim. of object
dim() # explore dimensions of two-dimensional obj.
nrow() # number of rows
ncol() # number of columns

head() # first few rows of data
tail() # last few rows of data
str() # explore structure of object

names() # names in the named vector - one dimension
rownames() # names of rows - two dimensions
colnames() # names of columns - two dimensions
```

Working directory

- Folder, where **all imports and exports are taking place** – enough to set once
- Makes data import and export easier
- Functions `setwd()` and `getwd()`
- 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")
```



	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

Libraries

- Libraries may be loaded using code

```
library("tm")

require("tm")
```

- Sometimes, libraries **conflict**
  - Packages may be unloaded when necessary

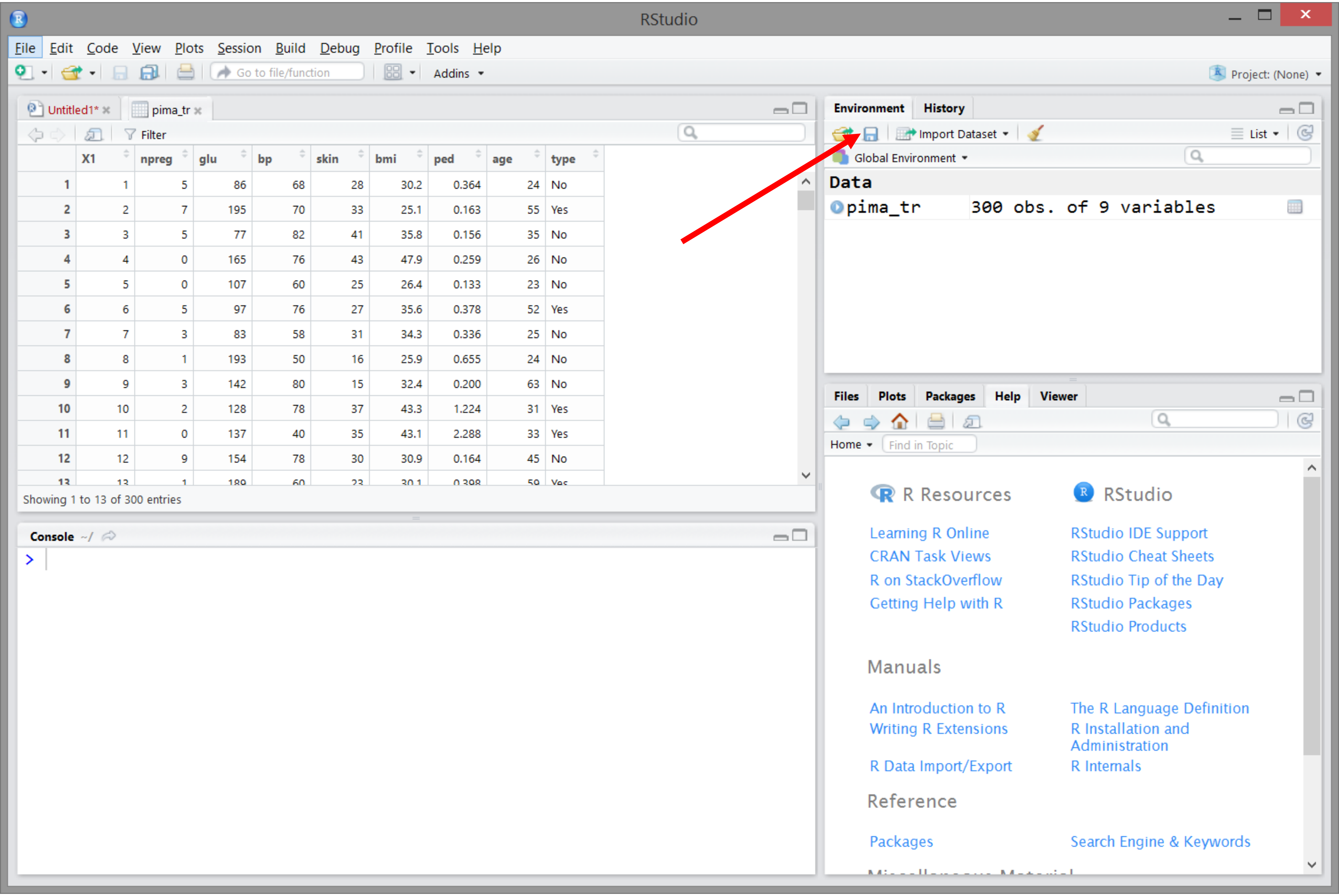
    ```
    detach("package:tm", unload = TRUE)
    ```

  - Instead of loading, individual functions may be called using name of package and ::

    ```
    tm::Corpus()
    ```

Data output

- Save entire workspace
  - Save all R objects you’ve created so far
  - Allows to return to work/backup current work
- Save particular object
  - Export data to tabular objects
  - CSV as most common format



	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
13	13	1	189	60	23	30.1	0.398	59	Yes

Showing 1 to 13 of 300 entries

```
>
```

Data

pima_tr 300 obs. of 9 variables

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

Miscellaneous Material

Exporting object – tabular

- Function `write.table()`
- **Name of file must be specified**
- Easy to import to Excel or other software

```
frequencies <- c(92,89,94,72,59)

write.table(frequencies,
            "frequencies.csv",
            sep = ",",
            row.names = FALSE,
            col.names = TRUE,
            fileEncoding = "UTF-8")
```

Topic modeling in R

Corpus

- We begin by working with “tm” package

- `getSources()` provides list of available sources
  - Files inside a directory – `DirSource()`
  - Text inside a vector – `VectorSource()`
  - Dataframe, XML, links to web-sites, …
- `Corpus()` creates a corpus object out of text sources

Corpus

- `DirSource()`
    - Read **all texts** in a directory
    - Attributes **help to qualify** which documents

Attribute	**Description**
`encoding`	Choose encoding of texts – usually useful to set to text value "UTF-8"
`pattern`	Look for file names which contain certain pattern. Useful to set to the most common text file extension "txt"
`recursive`	Logical attribute (TRUE/FALSE). If equals TRUE, files in all subdirectories will be included as well
`ignore.case`	Logical attribute (TRUE/FALSE). If equals TRUE, the case in the pattern matching will be ignored (e.g. if pattern equals to "txt", files with extension "TXT" will be included as well)

Corpus

```
require("tm")

my.dir <- "C:\\Users\\Lukas\\Desktop\\data\\"

directory.source <- DirSource(directory = my.dir,
                              encoding = "UTF-8",
                              ignore.case = T,
                              pattern = ".txt")

text.corpus <- Corpus(directory.source)
```

Corpus operations – functions

- Useful functions:
  - `removePunctuation()` – remove all punctuation
  - `removeWords()` – remove stopwords
  - `stripWhitespace()` – remove duplicate white space
  - `removeNumbers()` – remove all numbers
  - `stemDocument()` – stem document
  - `plainTextDocument()` – turn document into tm package's plain text format

Corpus operations

- `tm_map()` function allows to apply manipulations over the corpus data

```
edited.corpus <- text.corpus

edited.corpus <- tm_map(edited.corpus, removeNumbers)

edited.corpus <- tm_map(edited.corpus, removePunctuation)

edited.corpus <- tm_map(edited.corpus, stripWhitespace)

edited.corpus <- tm_map(edited.corpus,
                        removeWords,
                        stopwords("english"))
```

Document-term matrix

- Function `TermDocumentMatrix()`
  - Terms in rows
  - Documents/units in columns
- `DocumentTermMatrix()` creates **inverse** TDM
  - Documents/units in rows
  - Terms in columns
- Output is non-standard matrix object
  - If **matrix operations are needed**, it **must be converted** to basic matrix format with `as.matrix()` function
- **DTM is the input of the LDA**

Reduction of DTM

- Sparse terms add complexity, but **contribute little** to the analysis
  - May be dropped from the DTM
  - See for example Quinn et al. 2010
- `removeSparseTerms()`
  - Removes terms from the DTM
  - Sparsity is understood as relative measure
  - Sparsity is **percentage of documents** where term **is not present** (e.g. sparsity 0.99 represents term which does not occur in 99% of documents)

Term-document matrix

```
dtm <- DocumentTermMatrix(edited.corpus)

dtm <- removeSparseTerms(dtm,
                         sparse = 0.99)

dtm.matrixed <- as.matrix(dtm)
```

Running topic modeling – LDA

- Library “topicmodels” (Grün & Horink 2011)
- `LDA()` function
  - Uses **DTM only** (will **not accept TDM**)
  - Provides **two LDA topic models** –
    - LDA with Gibbs sampling (Griffiths & Styvers 2004)
    - LDA with VEM sampling algorithm (Blei, Ng & Jordan 2003)

LDA

- `LDA()` function attributes
    - DTM, number of topics and method of estimation has to be specified
    - Control attributes useful for more fine-grained control

Attribute	**Description**
`x`	DTM object – output from “tm” function
`k`	Number of topics
`method`	Method of sampling – either “VEM” or “Gibbs” is accepted
`control`	Additional attributes for estimation of the model

LDA control

- LDA `control` attribute requires **a list** of possible attributes

Attribute	**Description**
`verbose`	Positive number will make the function print continual information about the process of the estimation
`alpha`	Value of hyperparameter alpha – affecting “consistency” of topics
`iter`	Number of iterations in the Gibbs sampler (2000 by default)
`burnin`	Number of omitted iterations at the beginning (0 by default)
`thin`	Number of omitted in-between Gibbs iterations (equal to value of attribute `iter` by default)

LDA

```r
library(topicmodels)

n.topics <- 10

lda.parameters <- list(verbose = 1,
                       iter = 500,
                       thin = 300,
                       burnin = 1000,
                       alpha = 50/n.topics)

model <- LDA(x = dtm,
             k = n.topics,
             method = "Gibbs",
             control = lda.parameters)
```

Exploring results

- `terms()` function allows to explore chosen number of **most probable terms**
  - Attribute specifying number of terms is necessary
- `posterior()` function will **provide a list** of **two matrixes** containing probabilities
  - Probability of **terms** being in a topic
  - Probability of **documents** being drawn from topics

```
terms(model,10)

model.terms <- terms(model,10)
```

Exploring results

- To access results of `posterior()` function, we need to use **name of object, $ sign and name of sub-object**

```
model.posterior <- posterior(model)

model.posterior$topics

topic.doc.matrix <- model.posterior$topics

model.posterior$terms

topic.terms.matrix <- model.posterior$terms
```

Exporting results as tabular data

- Now time to use function `write.table()`
  - Exports either vector or data frame
  - Name of file **with extension** (txt, csv) must be specified
  - Easy to import to Excel or other software

```
write.table(topic.doc.matrix,
            "topic.doc.matrix.txt",
            sep = ",",
            row.names = TRUE,
            col.names = TRUE,
            fileEncoding = "UTF-8")
```

Other topic models

- Package “topicmodels”
  - Correlated topic model (Blei & Lefferty 2007) – model taking into accounts correlations between topics
- Package “lda”
  - Other implementation of LDA
  - Supervised LDA
  - Mixed-membership stochastic blockmodel
  - Relational Topic Model
- Package “mallet”, which connects Mallet software with R
- …