O čem píší čeští sociologové?

Česká sociologie očima jazykových modelů

Aleš Vomáčka

Katederní semestrálka 2024

To jsme já!

Katedra Sociologie...
a Výzkumný ústav Stem...
a Sociologický Ústav AV...

O čem to tu bude?

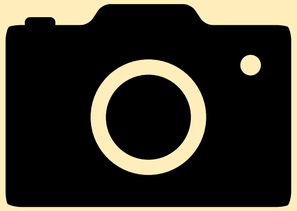
Strukturovaná data

0	1	0
1	0	1
0	1	0

Nestrukturovaná data

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse facilisis aliquam lorem, vel mattis ligula ullamcorper ac. Vivamus leo lorem, pulvinar ac euismod ut, feugiat iaculis sapien.

Aliquam mauris lacus, sodales quis scelerisque quis, rutrum sed mi. Sed accumsan varius tortor, sit amet volutpat mauris pellentesque id.



Cíle

1. Naučit se efektivně získat a zpracovat nestrukturovaný text

2. Zjistit, o čem že je ta česká sociologie

Kapitola 1.

Získání dat

Problém

Nestrukturovaná data většinou neexistují v jednoduše dostupné podobě

Testing the Psychometric Properties and Equivalence of the Czech Version of the Satisfaction with Life Scale (SWLS) using Confirmatory Factor Analysis, Item Response Theory, and Bayesian Modelling

Radka Hanzlová, Petra Raudenská

Sociologický ústav AV ČR, v. v. i., Praha

The Satisfaction with Life Scale (SWLS) is one of the most commonly used instruments for measuring life satisfaction. The aim of this study is to test the psychometric properties of the Czech version of the SWLS using Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) and to test its invariance between social groups in terms of gender, age, and education using Bayesian modelling on a representative sample of the Czech online population, as the scale has not yet been tested on representative data in the Czech Republic. The research sample consists of 960 respondents aged 18 to 69 years. The results confirmed that the psychometric properties of the Czech version of the SWLS are very good, but, at the same time, it is evident that the fifth item shows worse results than the other four items. In terms of dimensionality, CFA and IRT confirmed its modified single-factor structure with correlated residuals between the fourth and fifth items as the most appropriate. Testing for approximate measurement invariance using Bayesian modelling showed that the SWLS measures comparably between groups based on gender, age, and education. In conclusion, the Czech version of the SWLS is a suitable, verified, and reliable instrument for measuring the life satisfaction of Czech citizens.

Keywords: SWLS, Confirmatory Factor Analysis (CFA), Item Response Theory (IRT), Bayesian modelling BSEM, psychometrics, measurement invariance

Tohle chceme!

Pro každý článek... každého čísla... každého roku...

Tváří v tvář rutinní práci, sociologové tradičně spoléhají na MPS metodu

Mizerně Placené Stážisty

Jde ale to i jinak!

Co kdyby stážistou byl náš počítač?

WEB SCRAPING

Testing the Psychometric Properties and Equivalence of the Czech Version of the Satisfaction with Life Scale (SWLS) using Confirmatory Factor Analysis, Item Response Theory, and Bayesian Modelling

.title

Radka Hanzlová, Petra Raudenská

.author

Sociologický ústav AV ČR, v. v. i., Praha

.affiliation

The Satisfaction with Life Scale (SWLS) is one of the most commonly used instruments for measuring life satisfaction. The aim of this study is to test the psychometric properties of the Czech version of the SWLS using Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) and to test its invariance between social groups in terms of gender, age, and education using Bayesian modelling on a representative sample of the Czech online population, as the scale has not yet been tested on representative data in the Czech Republic. The research sample consists of 960 respondents aged 18 to 69 years. The results confirmed that the psychometric properties of the Czech version of the SWLS are very good, but, at the same time, it is evident that the fifth item shows worse results than the other four items. In terms of dimensionality, CFA and IRT confirmed its modified single-factor structure with correlated residuals between the fourth and fifth items as the most appropriate. Testing for approximate measurement invariance using Bayesian modelling showed that the SWLS measures comparably between groups based on gender, age, and education. In conclusion, the Czech version of the SWLS is a suitable, verified, and reliable instrument for measuring the life satisfaction of Czech citizens.

.abstract

Keywords: SWLS, Confirmatory Factor Analysis (CFA), Item Response Theory (IRT), Bayesian modelling BSEM, psychometrics, measurement invariance

.keywords

issue	article_link	title	authors	affiliation	annotation	tags
1	/artkey/csr-20240	Moderní postupy v modelování kvantitativních sociálněvědních dat	Petr Soukup ORCID...	Fakulta sociálních věd, Univerzita Karlova, Praha		
1	/artkey/csr-20240	Exponenciální modely náhodných grafů: modelování relačních mechanismů na případu sítě organizací zapojených v českém uhelném sektoru	Tomáš Diviák ORCID...1, Petr Ocelík ORCID...2	1 Department of Criminology a Mitchell Centre for Social Network Analysis, University of Manchester;2 Fakulta sociálních studií, Masarykova univerzita, Brno	This study provides the first comprehensive introduction to exponential random graph models (ERGM) in the Czech academic literature. In it we apply ERGM to a network of 68 organisations involved in the Czech coal policy subsystem...	Klíčová slova: social network analysis, exponential random graph models, political networks, social mechanisms, statistical models
1	/artkey/csr-20240	Testování psychometrických vlastností a ekvivalence české verze škály spokojenosti se životem (SWLS) pomocí metod konfirmační faktorové analýzy, teorie odpovědi na položku a bayesovského modelování	Radka Hanzlová ORCID..., Petra Raudenská ORCID...	Sociologický ústav AV ČR, v. v. i., Praha	The Satisfaction with Life Scale (SWLS) is one of the most commonly used instruments for measuring life satisfaction. The aim of this study is to test the psychometric properties of the Czech version of the SWLS...	Klíčová slova: SWLS, Confirmatory Factor Analysis (CFA), Item Response Theory (IRT), Bayesian modelling BSEM, psychometrics, measurement invariance
1	/artkey/csr-20240	Využití víceúrovňových modelů při analýze kontextuálních efektů míry ekonomické aktivity na podporu přerozdělování v komparativních longitudinálních datech	Ivan Petrúšek ORCID...	Sociologický ústav AV ČR, v. v. i., Praha	This article studies the links between a country's labour force participation rate and attitudes towards income redistribution. The article also demonstrates how to specify a multilevel models...	Klíčová slova: multilevel models, contextual effects, redistribution support, random effects, centring variables, comparative longitudinal data

Od roku 2009 vydal ČSR 1353 dokumentů

Z nich bylo 414 (30%) vědecké studie

Vědecké studie publikovalo 486 autorů/autorek

Průměrný autor/ka publikoval/a 1.5 článku

Jedna studia má v průměru 1.7 autorů/ek

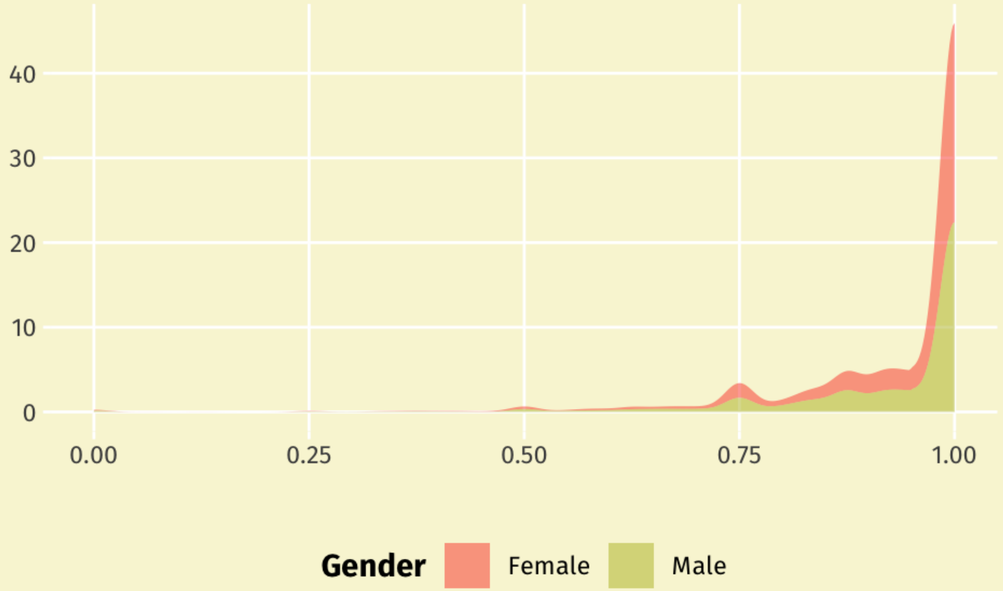
Pod nejpočetnější studí je podepsáno 7 osob

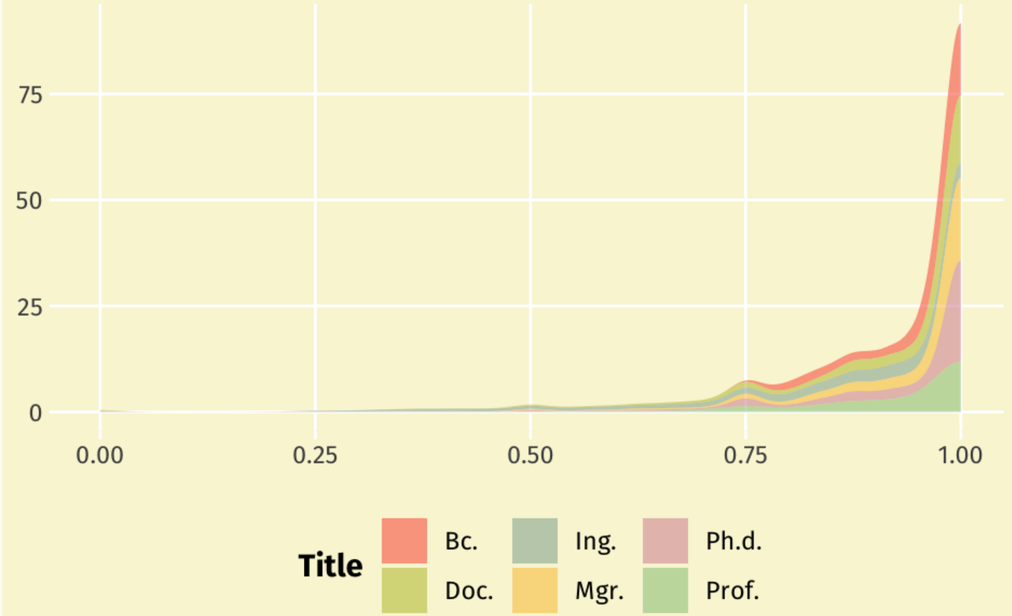
Web scraping je cool!

Případová studie: Studentská výuky hodnocení na FF UK

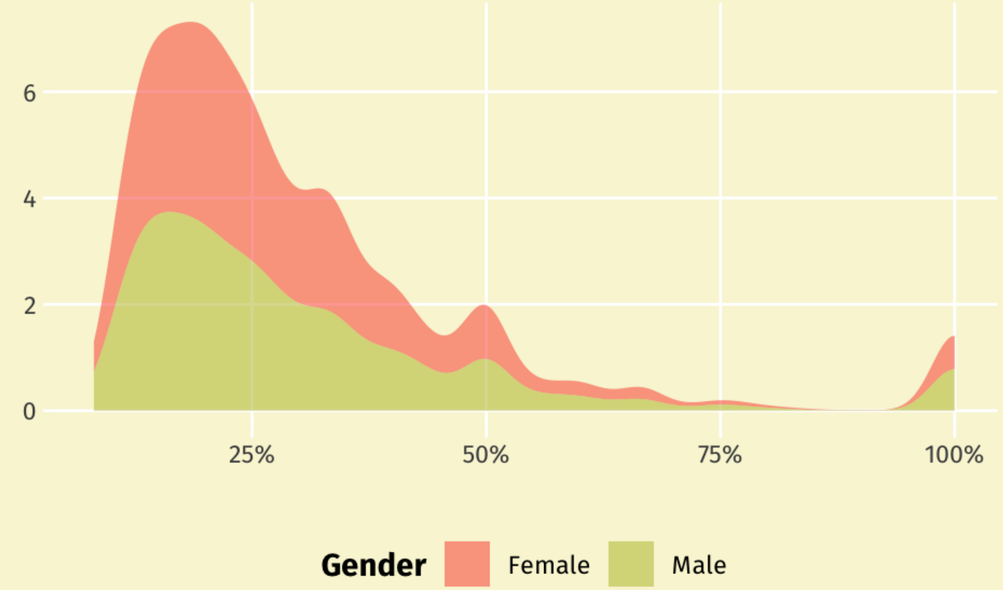
Jazykové centrum - francouzština	9	30 / 122	24.59%
Jazykové centrum - italština	1	0 / 17	0%
Jazykové centrum - latina	12	45 / 158	28.48%
Jazykové centrum - němčina	16	52 / 267	19.48%
Jazykové centrum - ruština	7	14 / 63	22.22%
Jazykové centrum - španělština	7	21 / 90	23.33%
Katedra Blízkého východu	93	188 / 904	20.8%
Katedra PVH a archivního studia	82	96 / 738	13.01%
Katedra andragogiky a personálního řízení	54	196 / 965	20.31%
Katedra divadelní vědy	32	52 / 307	16.94%
Katedra estetiky	21	31 / 283	10.95%
Katedra filmových studií	47	57 / 609	9.36%
Katedra jihoslovanských a balkanistických studií	98	43 / 369	11.65%
Katedra logiky	19	14 / 145	9.66%
Katedra pedagogiky	114	120 / 2170	5.53%
Katedra psychologie	106	535 / 2493	21.46%
Katedra sinologie	42	100 / 421	23.75%
Katedra sociologie	49	257 / 1125	22.84%
Katedra sociální práce	41	128 / 796	16.08%

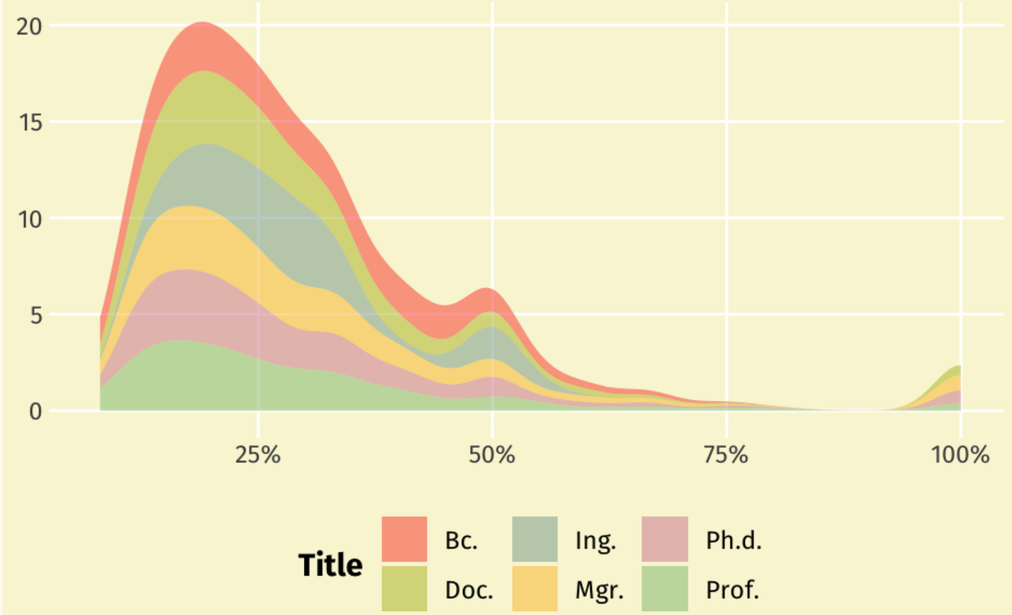
Teacher Rating (Higher is Better)





Response Rate





Bibliometrie je zajímavá, ale sociologie je o zkoumání sociálních skupin.

Třeba těch genderových.

Ale jak zjistíme gender/pohlaví autorstva?

Kapitola 2.

API

Testing the Psychometric Properties and Equivalence of the Czech Version of the Satisfaction with Life Scale (SWLS) using Confirmatory Factor Analysis, Item Response Theory, and Bayesian Modelling

Radka Hanzlová, Petra Raudenská

Sociologický ústav AV ČR, v. v. i., Praha

Máme jména, ale ne pohlaví...

The Satisfaction with Life Scale (SWLS) is one of the most commonly used instruments for measuring life satisfaction. The aim of this study is to test the psychometric properties of the Czech version of the SWLS using Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) and to test its invariance between social groups in terms of gender, age, and education using Bayesian modelling on a representative sample of the Czech online population, as the scale has not yet been tested on representative data in the Czech Republic. The research sample consists of 960 respondents aged 18 to 69 years. The results confirmed that the psychometric properties of the Czech version of the SWLS are very good, but, at the same time, it is evident that the fifth item shows worse results than the other four items. In terms of dimensionality, CFA and IRT confirmed its modified single-factor structure with correlated residuals between the fourth and fifth items as the most appropriate. Testing for approximate measurement invariance using Bayesian modelling showed that the SWLS measures comparably between groups based on gender, age, and education. In conclusion, the Czech version of the SWLS is a suitable, verified, and reliable instrument for measuring the life satisfaction of Czech citizens.

Keywords: SWLS, Confirmatory Factor Analysis (CFA), Item Response Theory (IRT), Bayesian modelling BSEM, psychometrics, measurement invariance

I tento problém jde řešit MPS metodou.

I tady ale existují efektivnější řešení.

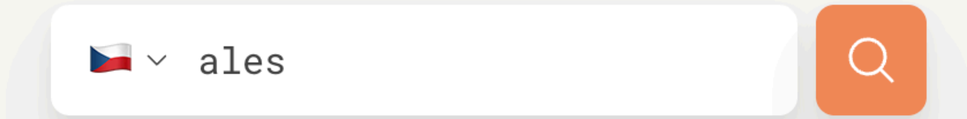
Application Progamming Interface

Standardizovaný způsob, jak komunikovat s online službami.

Službami pro získávání dat z sociálních sítí, klasifikaci obrázku a mnoho dalšího

genderize.io

Check the Gender of a Name



ales in **Czechia** is **male** with **99%** certainty

Query
https://api.genderize.io?name=ales&country_id=CS

Response


```
{
    "name": "ales",
    "gender": "male",
    "country_id": "CS",
    "probability": 0.99
}
```

U 54% článků byl prvním autorem muž

Články, jejichž prvním autorem muž, mají v průměru 1.8 autorů/autorek (median = 2)

Články, jejichž prvním autorem žena, mají v průměru 1.7 autorů/autorek (median = 1)

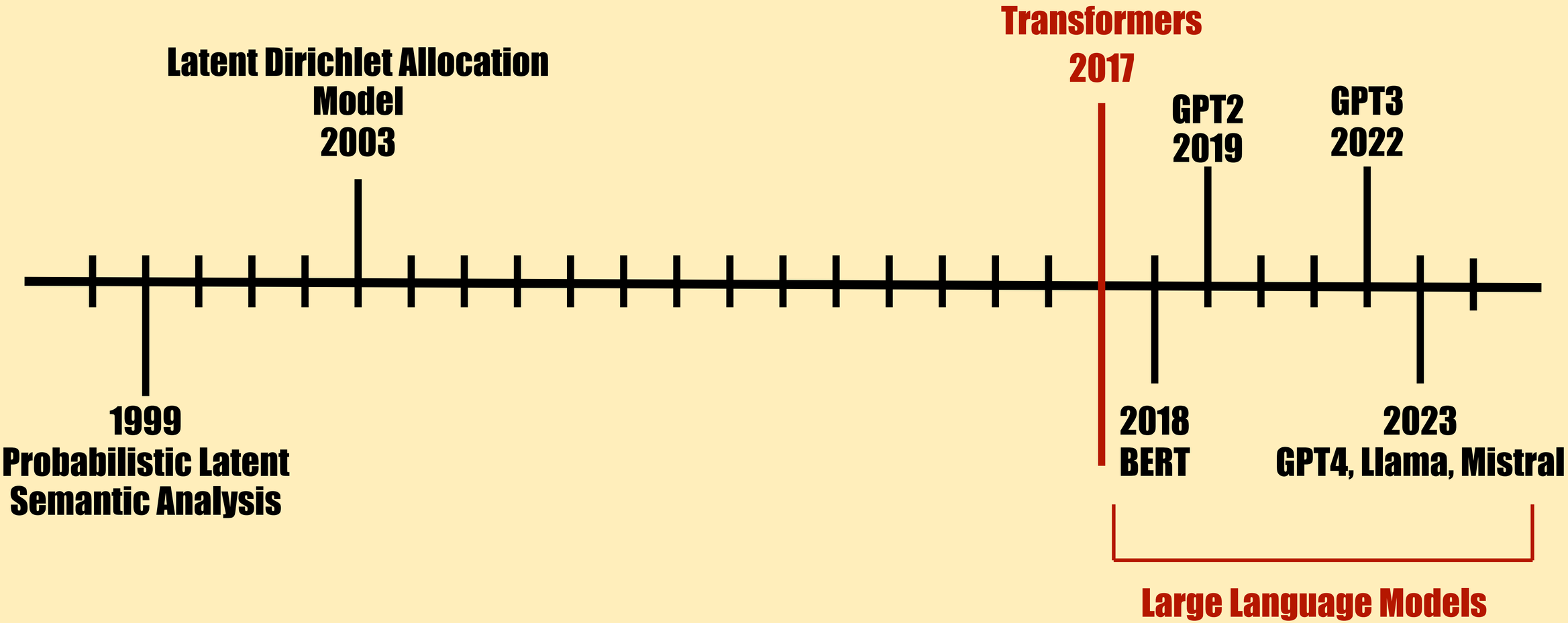
Webové služby cool!

Víme, kolik muži a ženy publikují, jak často a s kým.

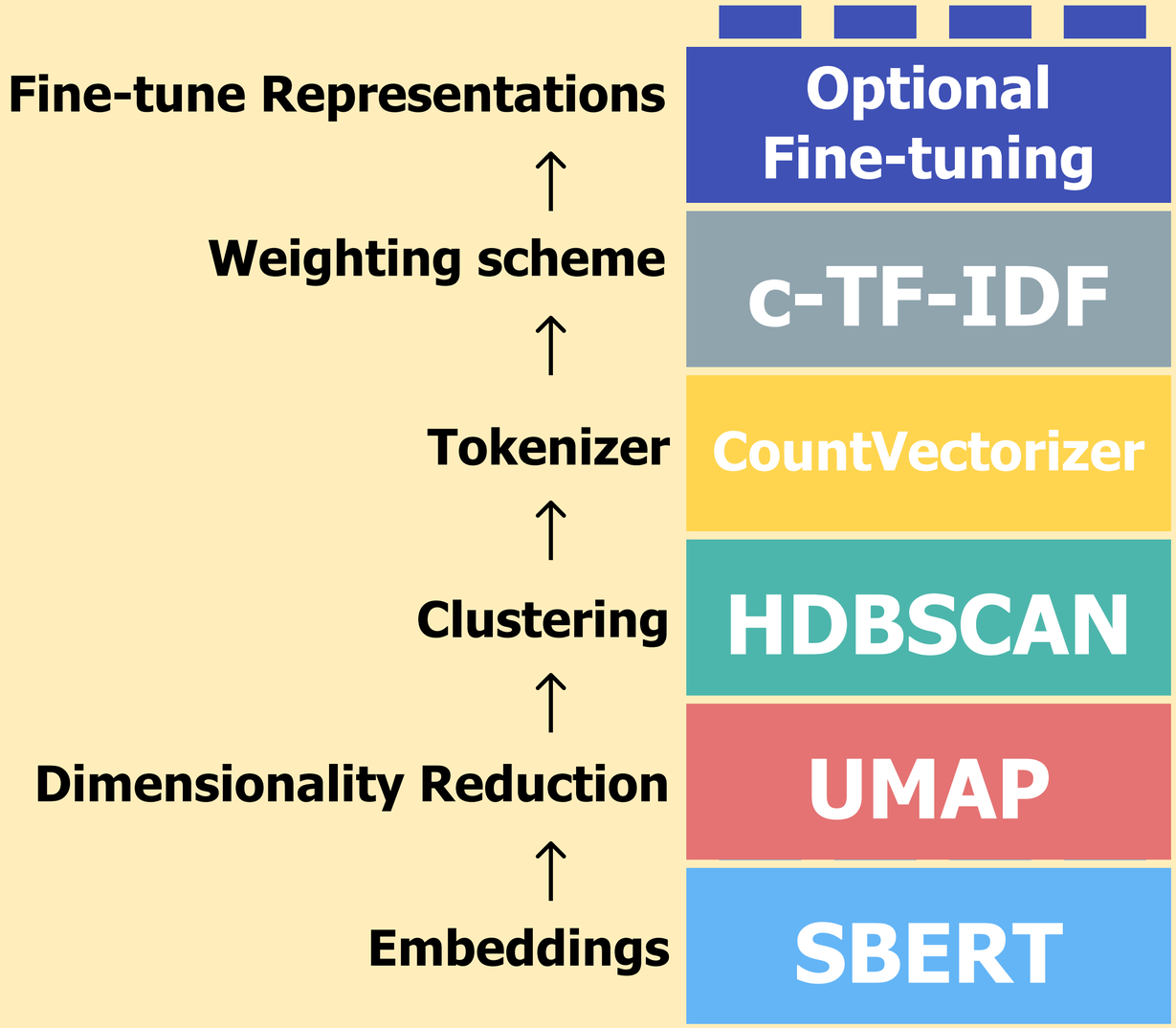
Ale liší se sociologové a sociložky v tom, o čem píší?

Kapitola 3.

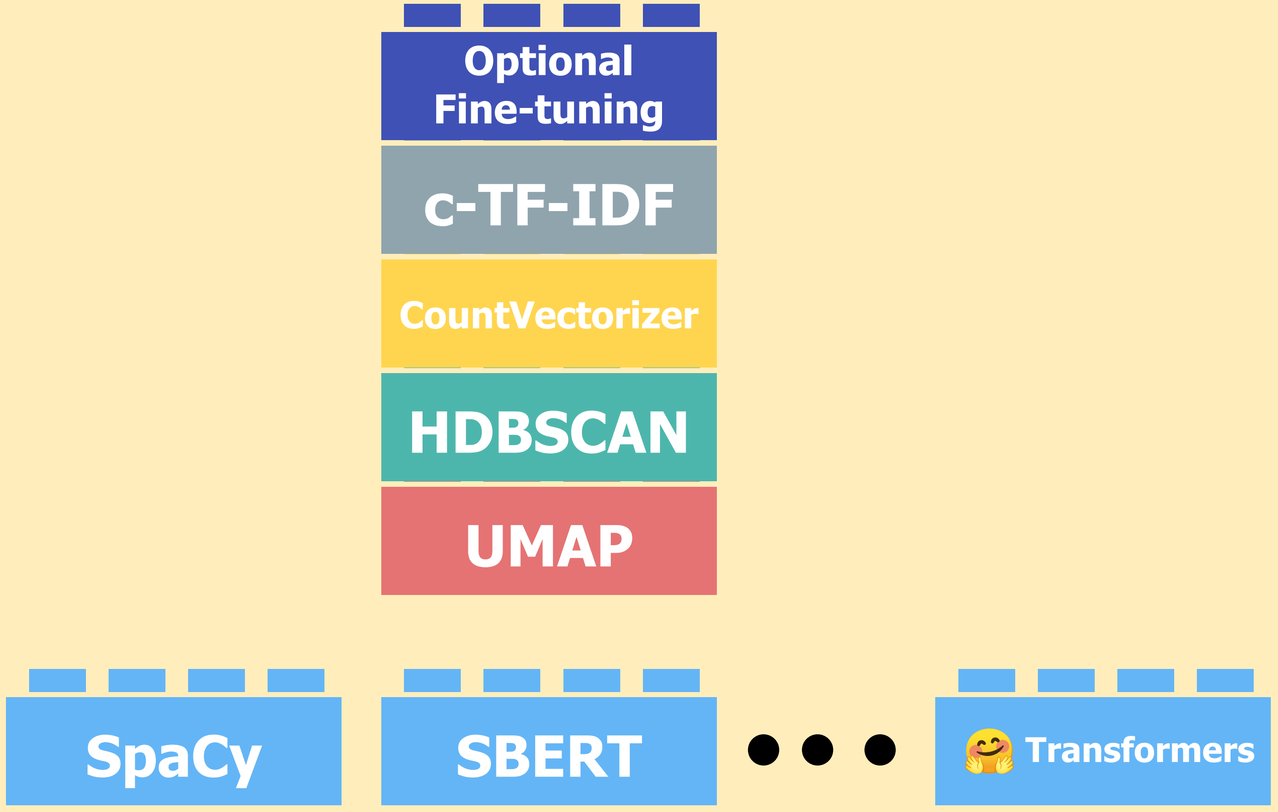
Topic Modeling


Transformers
2017
Latent Dirichlet Allocation
Model
2003
GPT2
2019
GPT3
2022
1999
Probabilistic Latent
Semantic Analysis
2018
BERT
2023
GPT4, Llama, Mistral
Large Language Models

Topic modeling using BERTopic



Embeddings

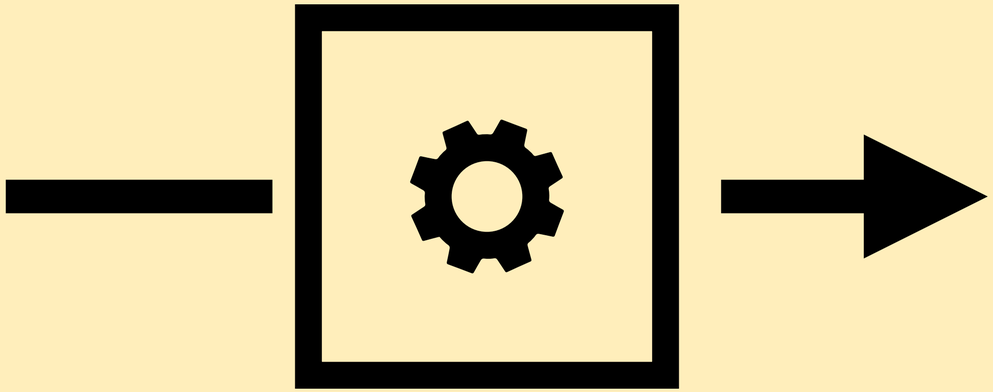

Optional
Fine-tuning
c-TF-IDF
CountVectorizer
HDBSCAN
UMAP
SpaCy
SBERT
Transformers

Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse facilisis aliquam lorem, vel mattis ligula ullamcorper ac. Vivamus leo lorem, pulvinar ac euismod ut, feugiat iaculis sapien.

Aliquam mauris lacus, sodales quis scelerisque quis, rutrum sed mi. Sed accumsan varius tortor, sit amet volutpat mauris pellentesque id.

Transformer



Embeddings

0	1	0
1	0	1
0	1	0

Embeddings kódují významovou blízkost slov

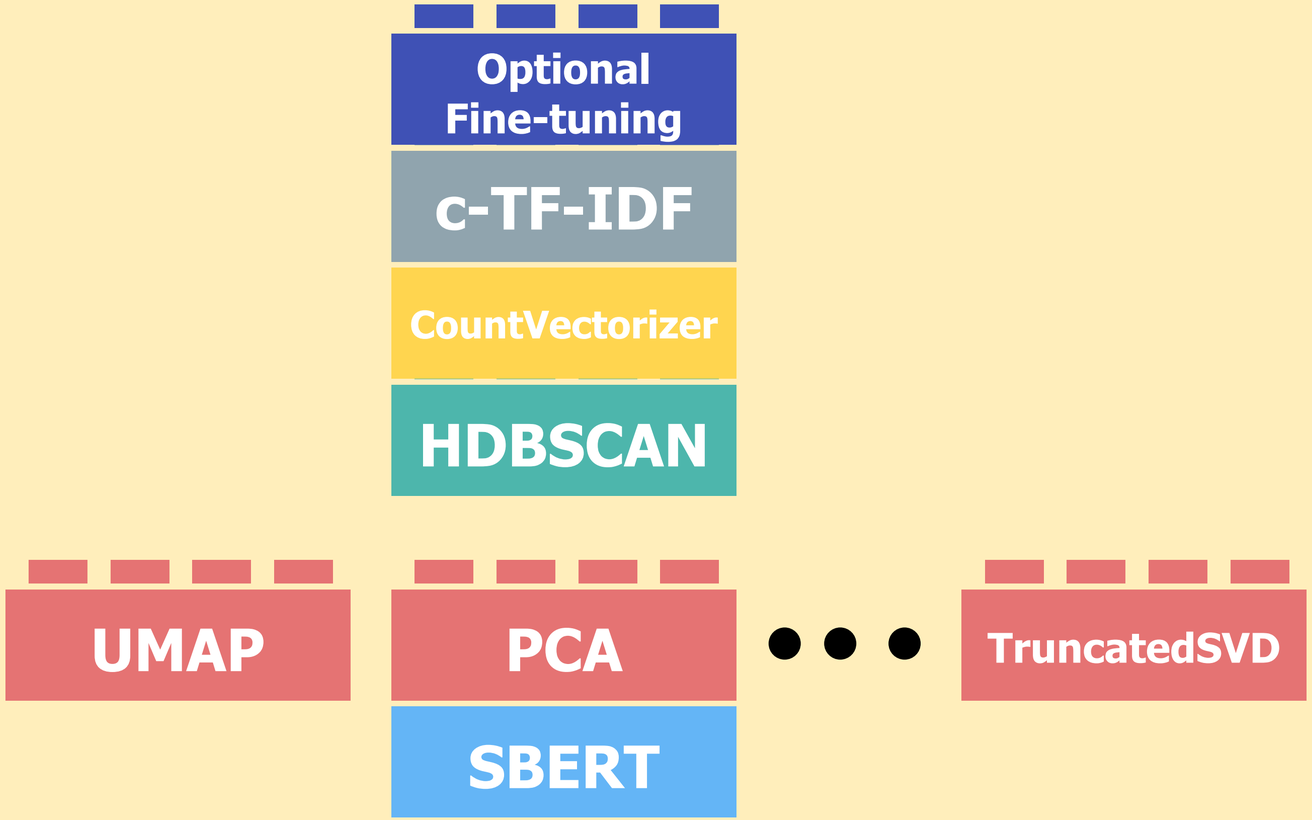
King	-	Man	+	Woman	=	Queen
2 4 6 8 5 2 9 6 1 7 5 8 8 4 0 3		6 5 4 0 6 4 6 1 2 3 5 8 1 3 9 2		0 2 4 1 2 2 6 2 3 8 5 3 1 0 7 6		-4 1 6 9 1 0 9 7 2 12 5 3 8 1 -2 7

Výběr modelu na základě dvou hlavních paramaterů

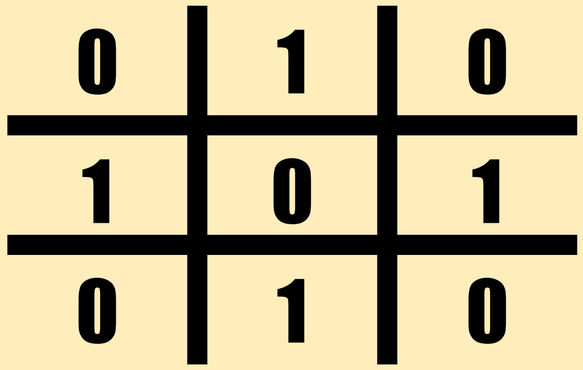
Kvality embeddings a **počtu tokenů**, které model dokáže zpracovat

1 token ~ 0.6 slova

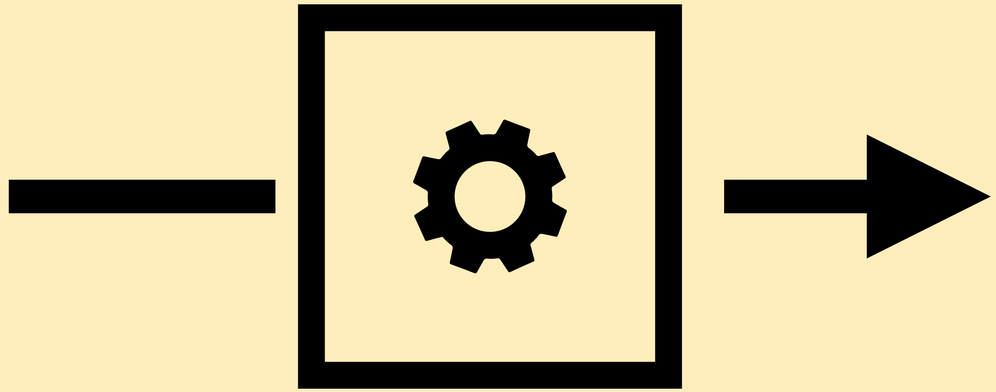
Dimensionality Reduction



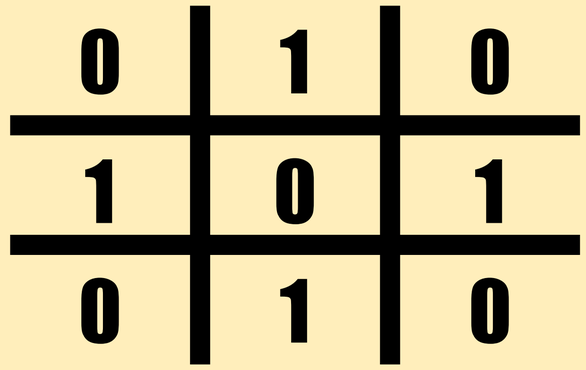
Embeddings



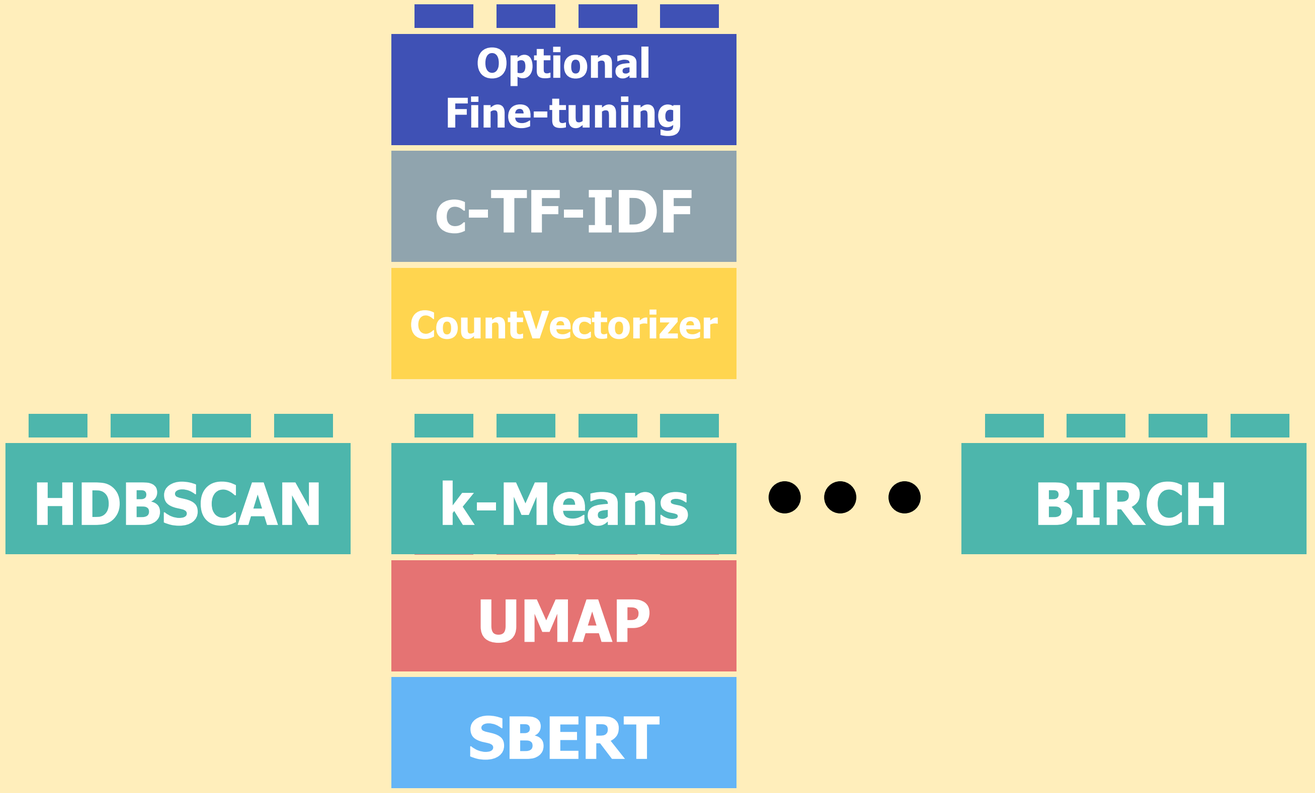
PCA/UMAP



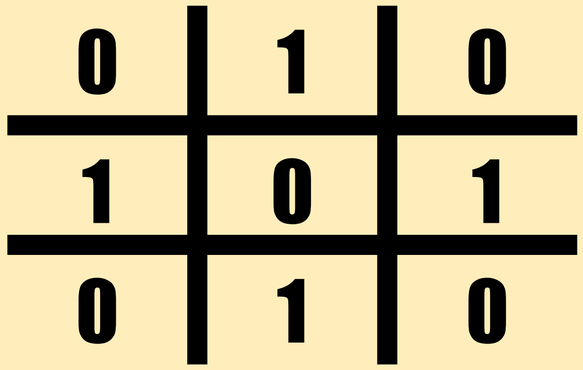
Embeddings s menším počtem dimenzí



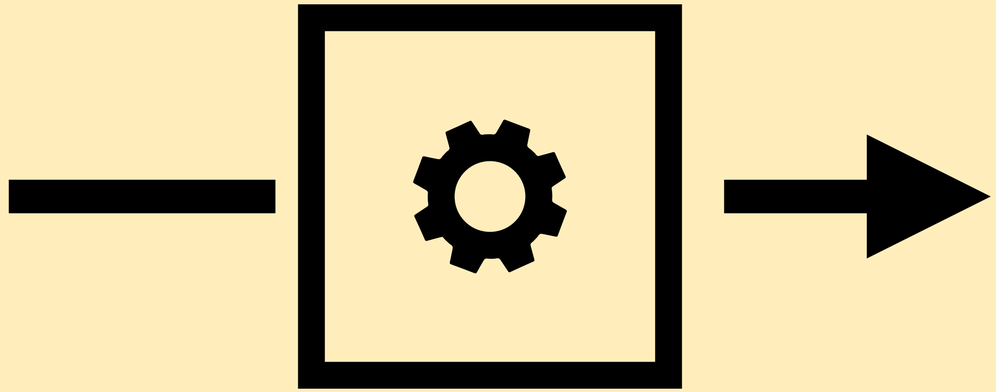
Clustering


Optional
Fine-tuning
c-TF-IDF
CountVectorizer
HDBSCAN
k-Means
BIRCH
UMAP
SBERT

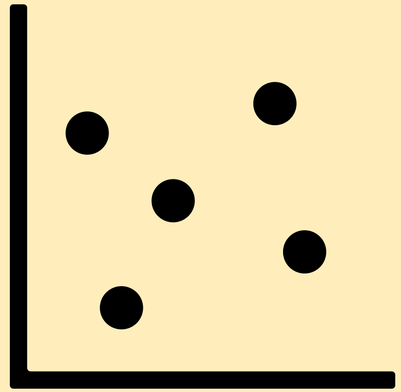
Původní Embeddings



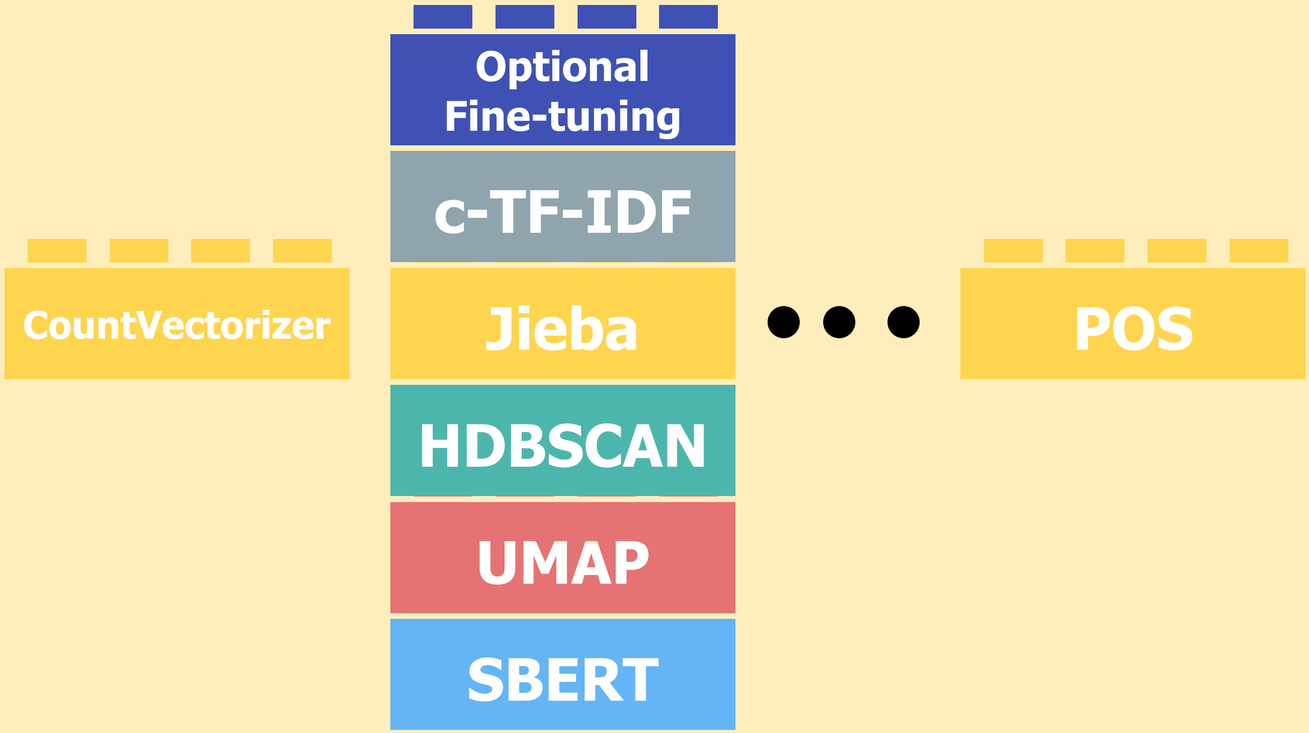
KMEANS/ HDBSCAN



Dokumenty rozřazené do klastrů



Vectorizers



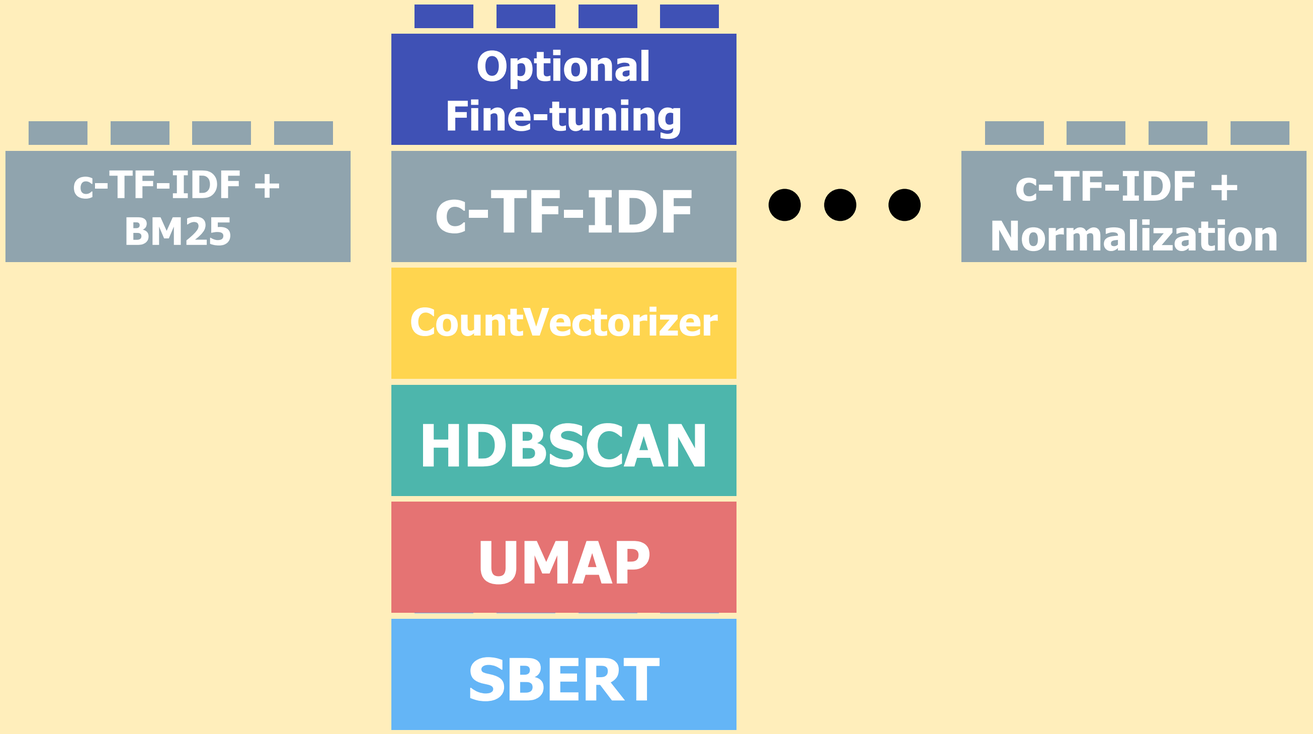
Jaká slova vybrat jako reprezentativní pro každé téma?

Stopwords (“článek”, “studie”, sociologický”)

Ngramy (“social” a “capital” vs “social capital”)

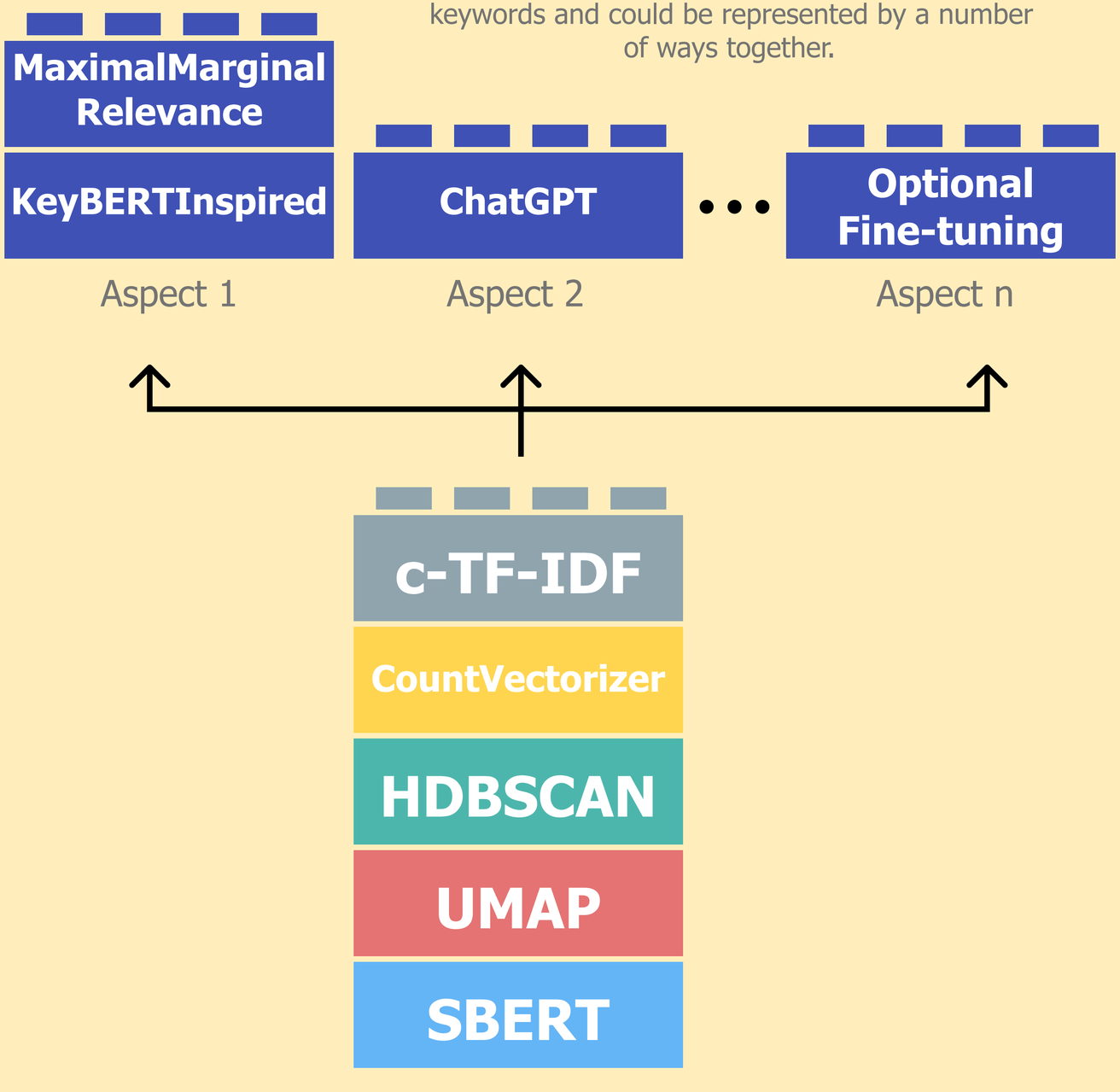
Minimální výskyt slov

Identifying key words

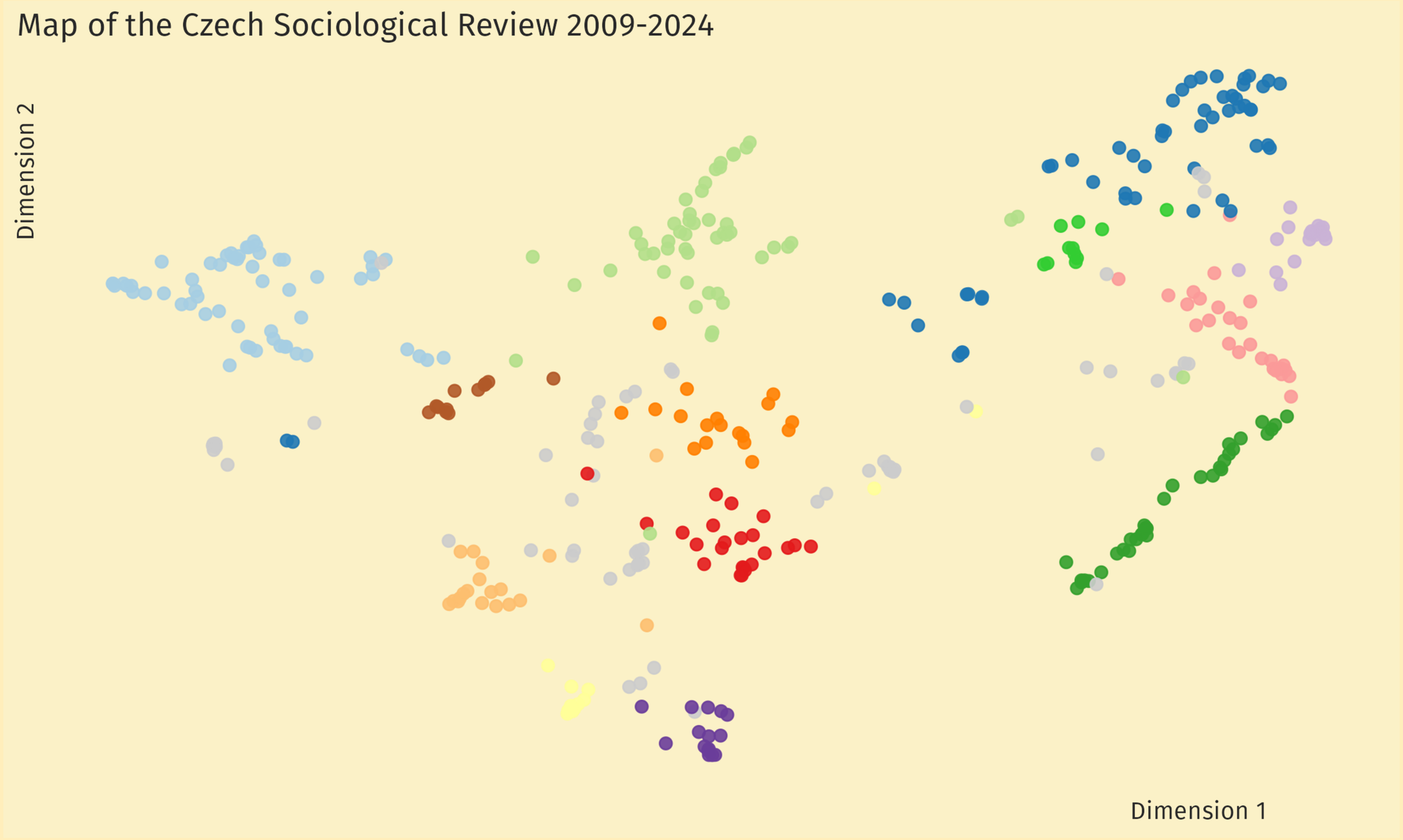


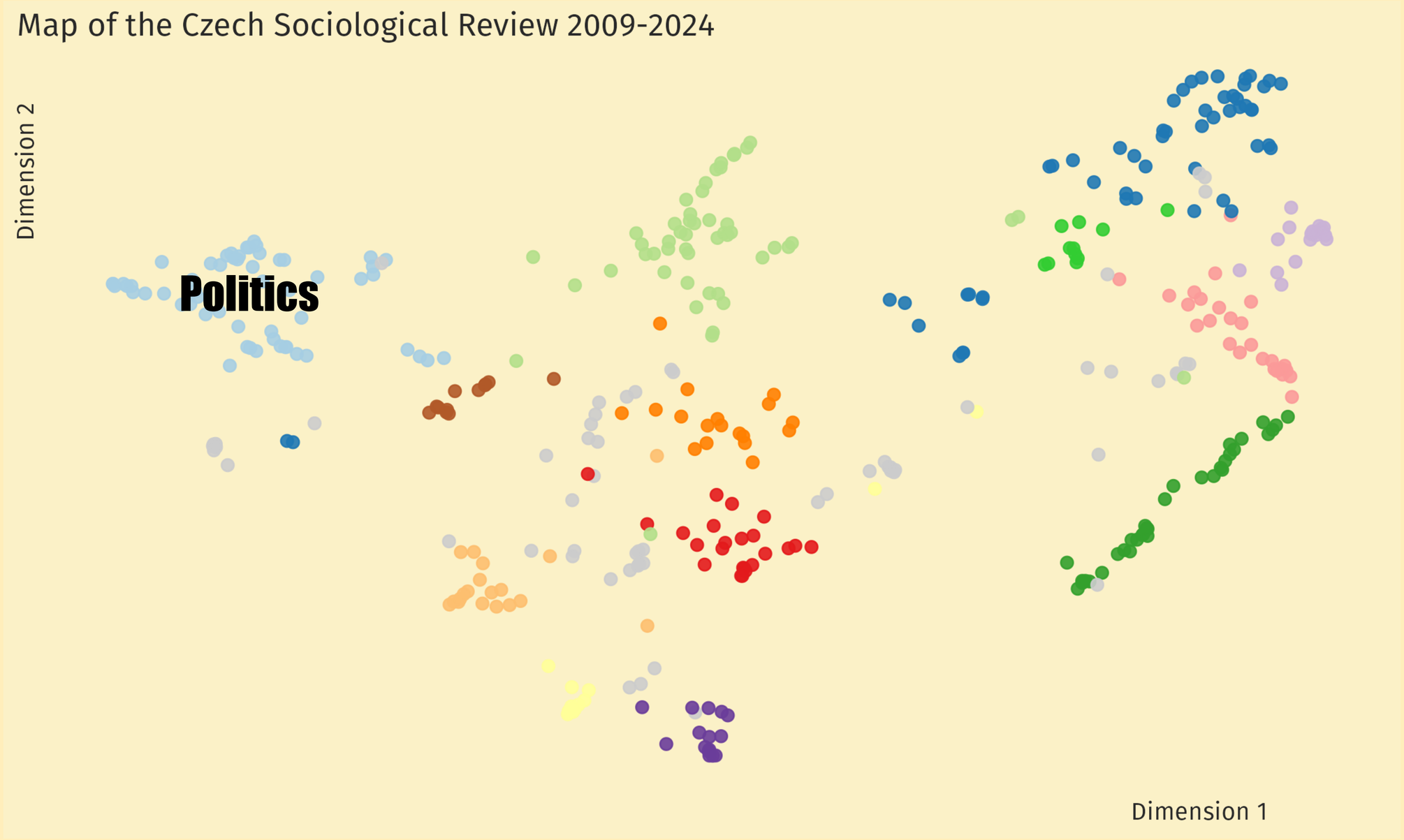
Fine-tuning (optional)

Create multiple topic representations, or aspects, simultaneously. Topics are more than just keywords and could be represented by a number of ways together.

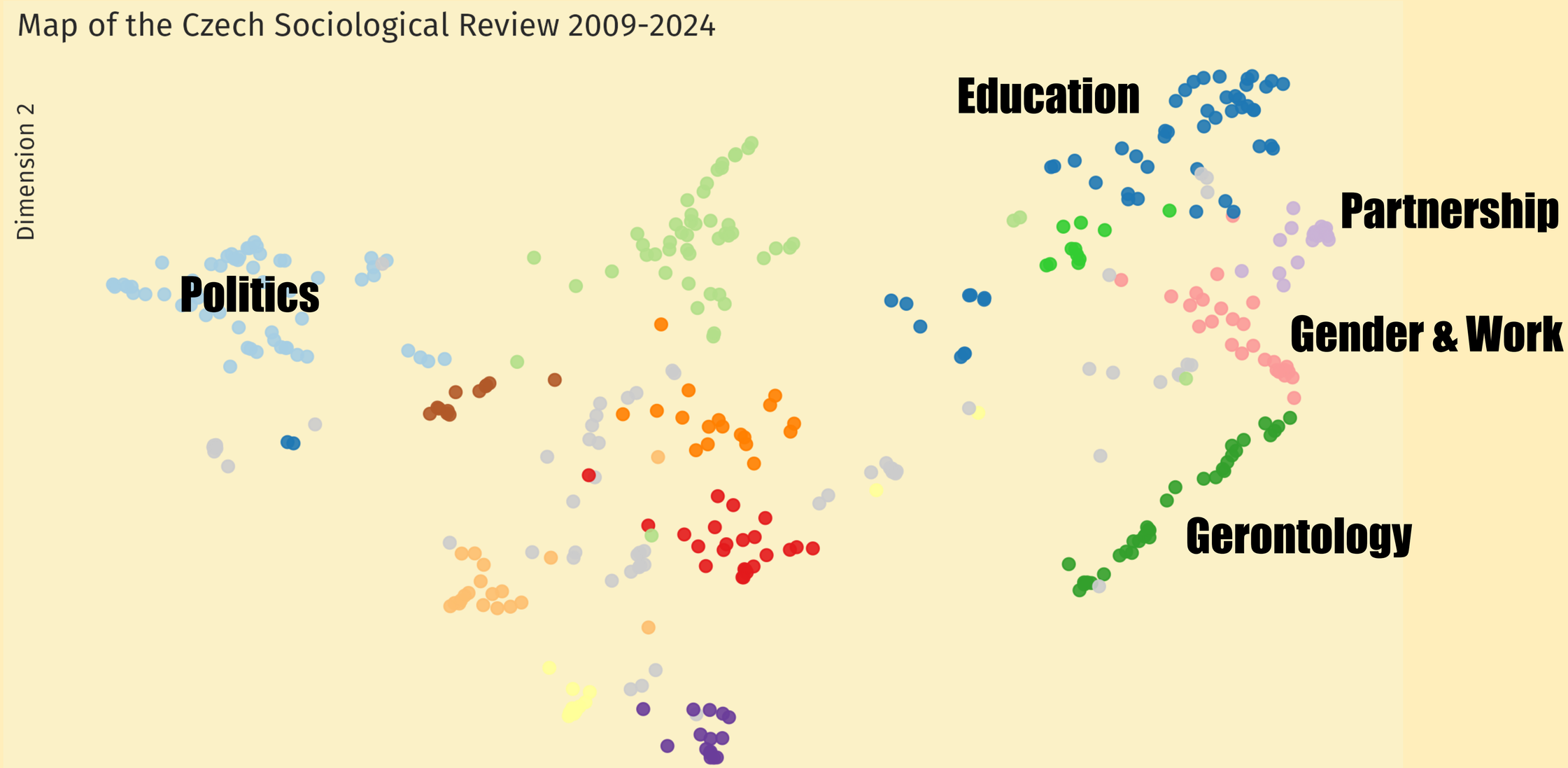


A co jsme zjistili?


Map of the Czech Sociological Review 2009-2024
Dimension 2
Dimension 1



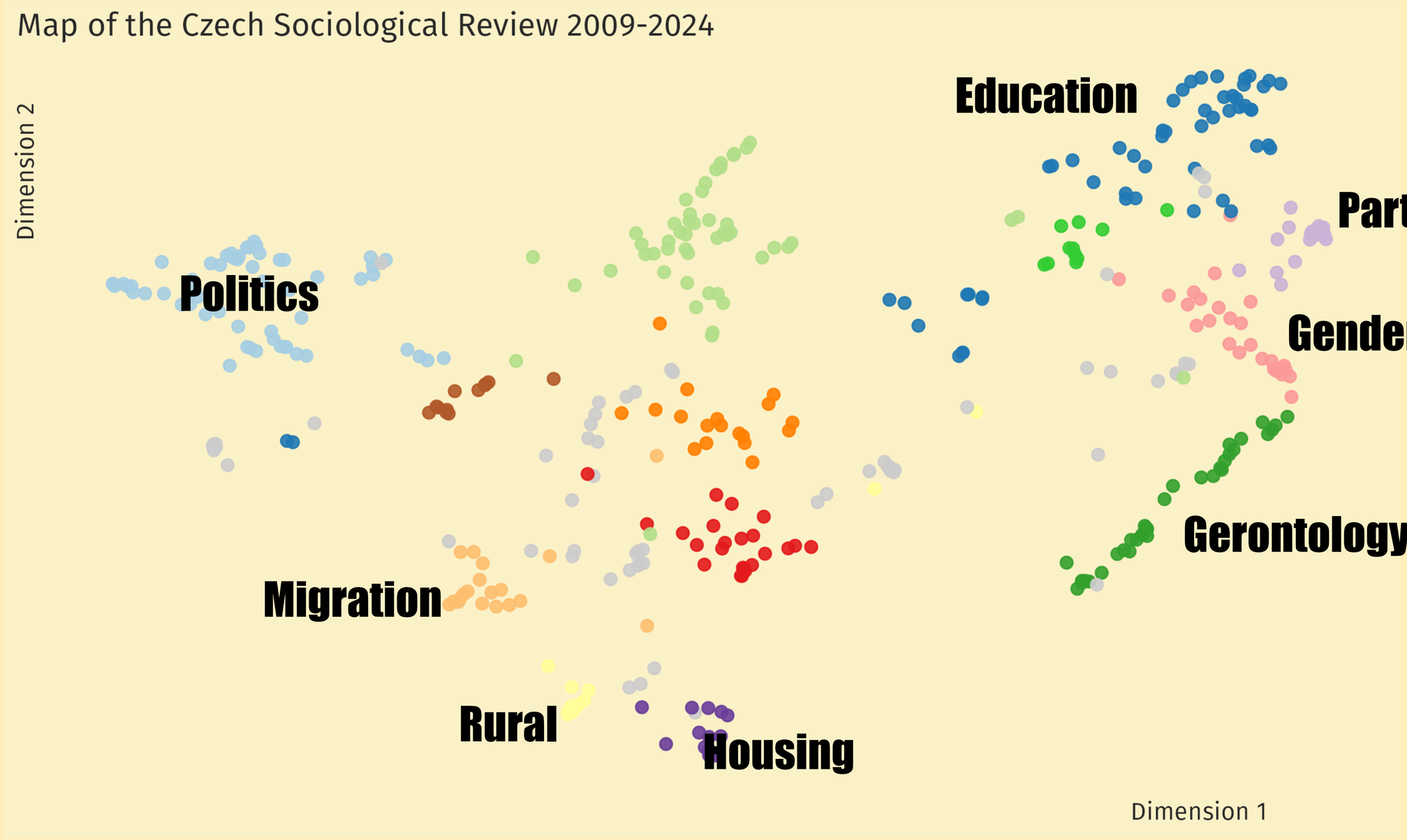
Strany na ústupu, lídři na vzestupu? Personalizace volebního chování v České republice



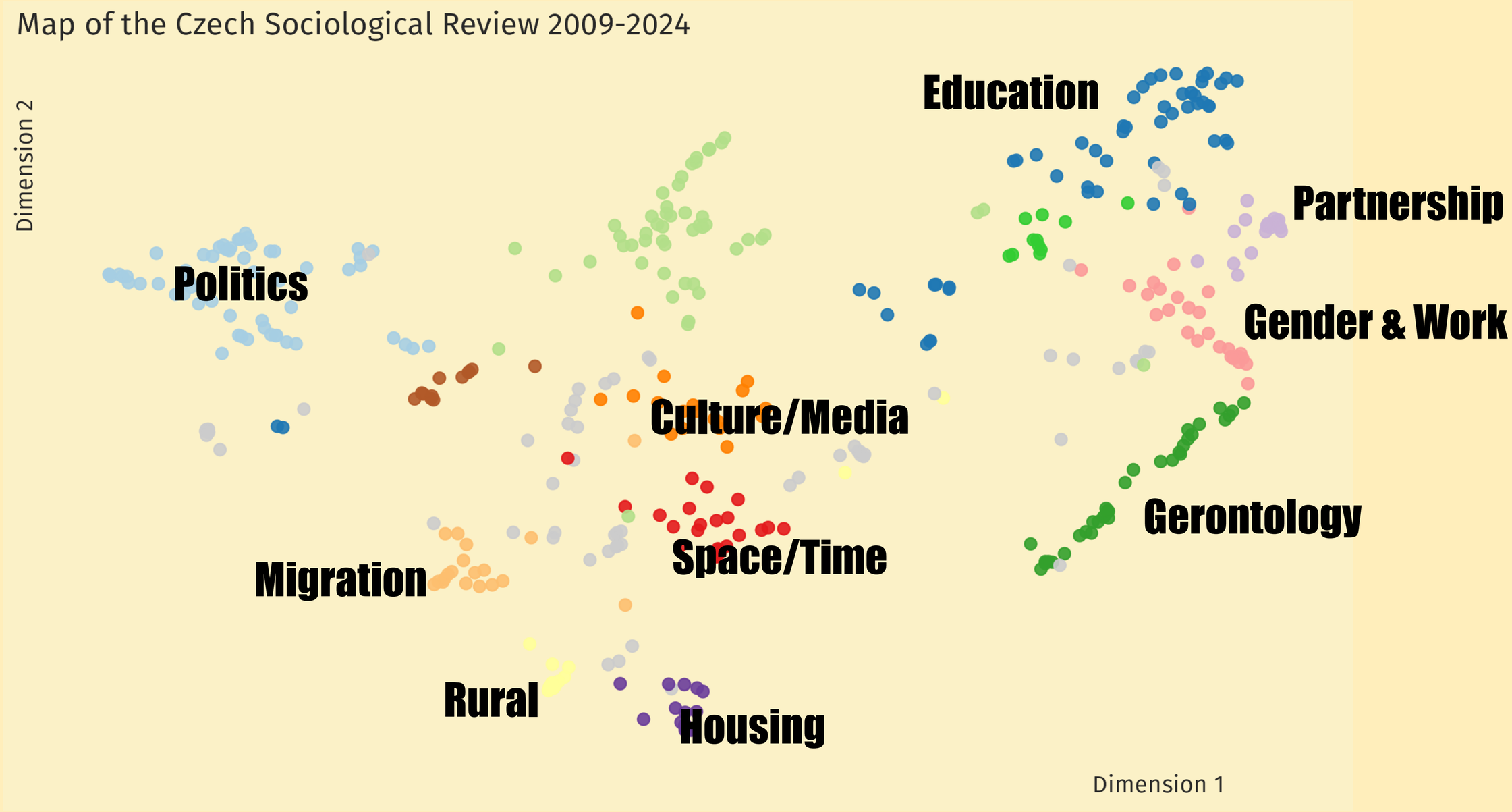
Neplodnost jeho a neplodnost její: Genderové aspekty asistované reprodukce

Sourozenectví v pozdním věku: Příspěvek k teoretické diskusi

Vliv postojů učitelů na výsledky žáků

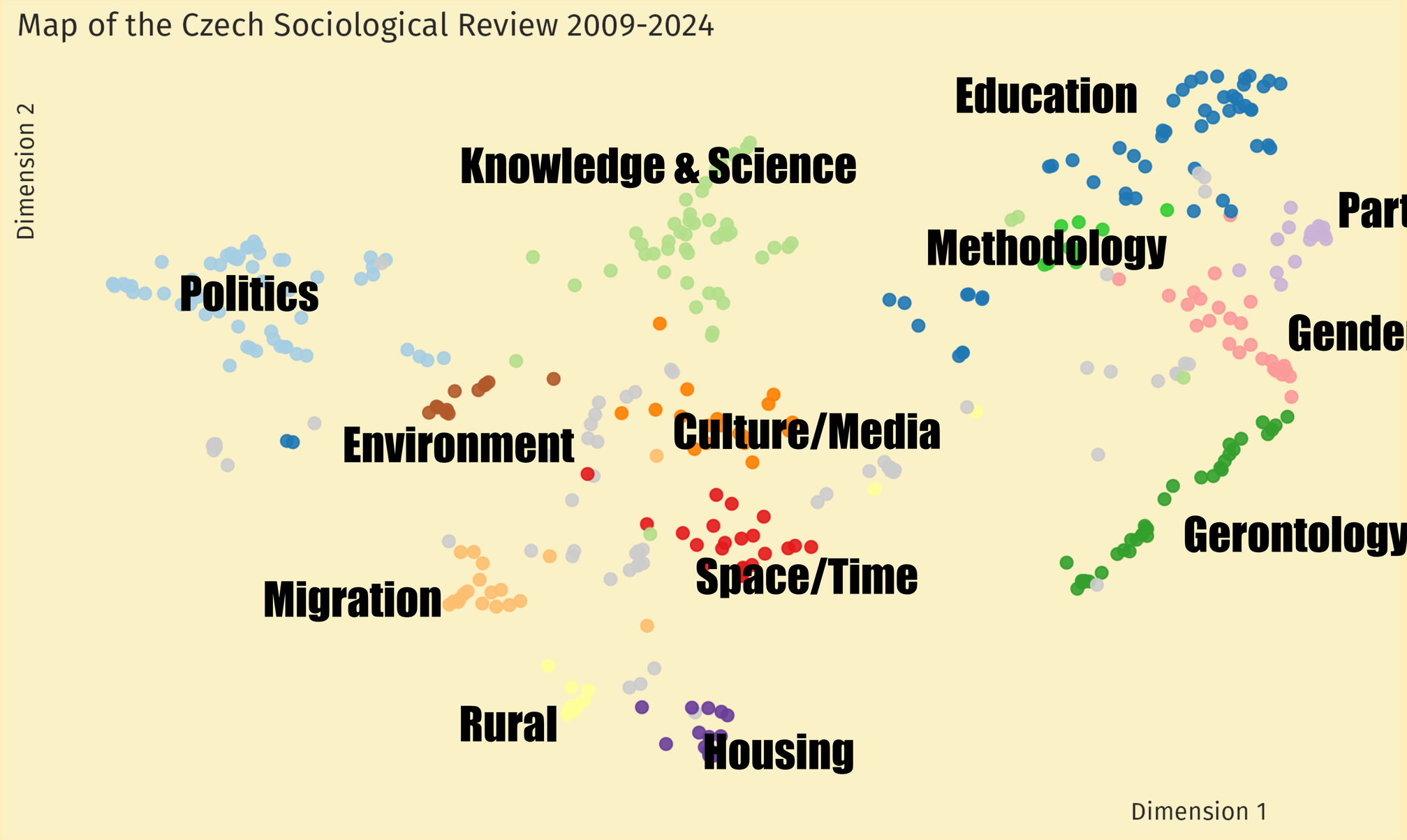


Private Rental Housing in the Czech Republic: Growth and...?
Endogenní rozvojové potenciály malých venkovských obcí
Transnacionalismus a stálost návratu v arménské návratové migraci



Discourses of Thrift and Consumer Reasonability in Czech State-Socialist Society

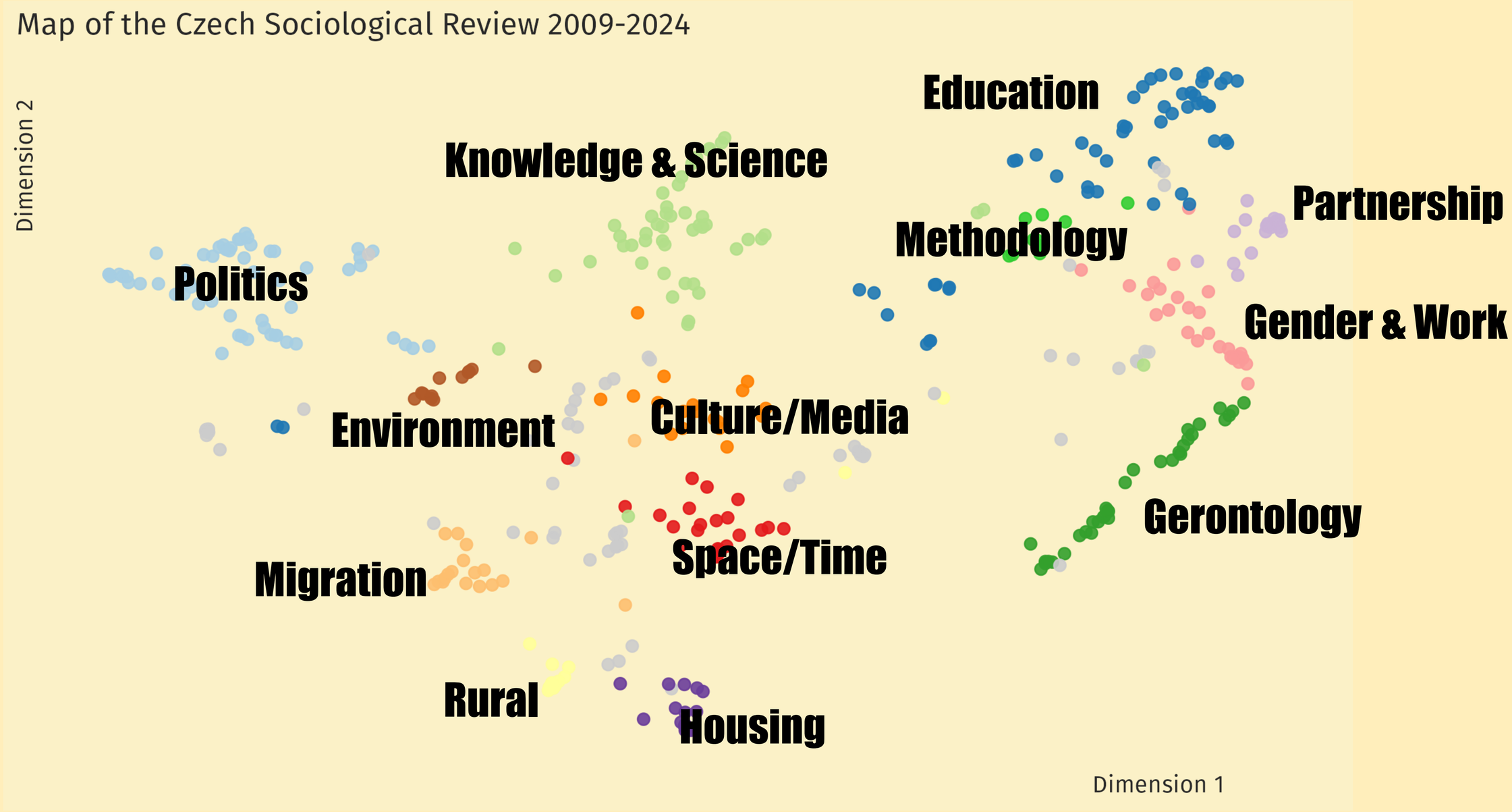
"Předevčírem, nebo kdy to bylo?": Temporalita třídy nejchudších



K rekonstrukci "kapitalismu" u autorů otevřeného marxismu

Measurement Invariance of the SQWLi Instrument Over Time

Environmentálně orientované motivace a potenciál zklamání


Map of the Czech Sociological Review 2009-2024
Dimension 2
Education
Knowledge & Science
Partnership
Methodology
Politics
Gender & Work
Environment
Culture/Media
Gerontology
Space/Time
Migration
Rural
Housing
Dimension 1

Nic není dokonalé

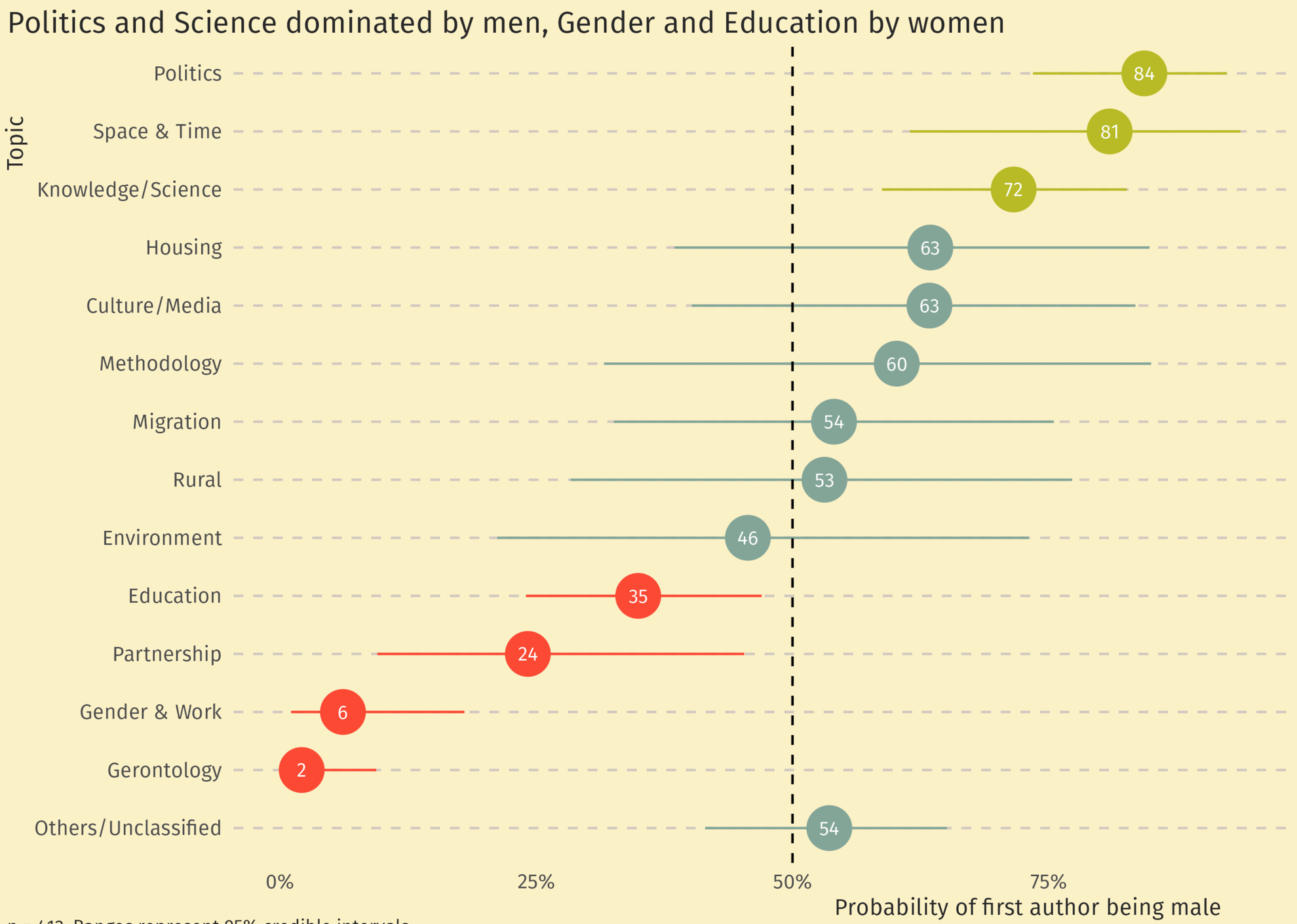
Trends in Divorce Acceptance and Its Correlates across European Countries

Education?

Exponenciální modely náhodných grafů: modelování relačních mechanismů na případu sítě organizací zapojených v českém uhelném sektoru

Outlier/Unclassified?

Témata a Gender



n = 412. Ranges represent 95% credible intervals

Popularita témat v čase

Tématická heterogenita autorů autorů

Síťová analýza autorů

A další...

“Konfirmativní” analýza

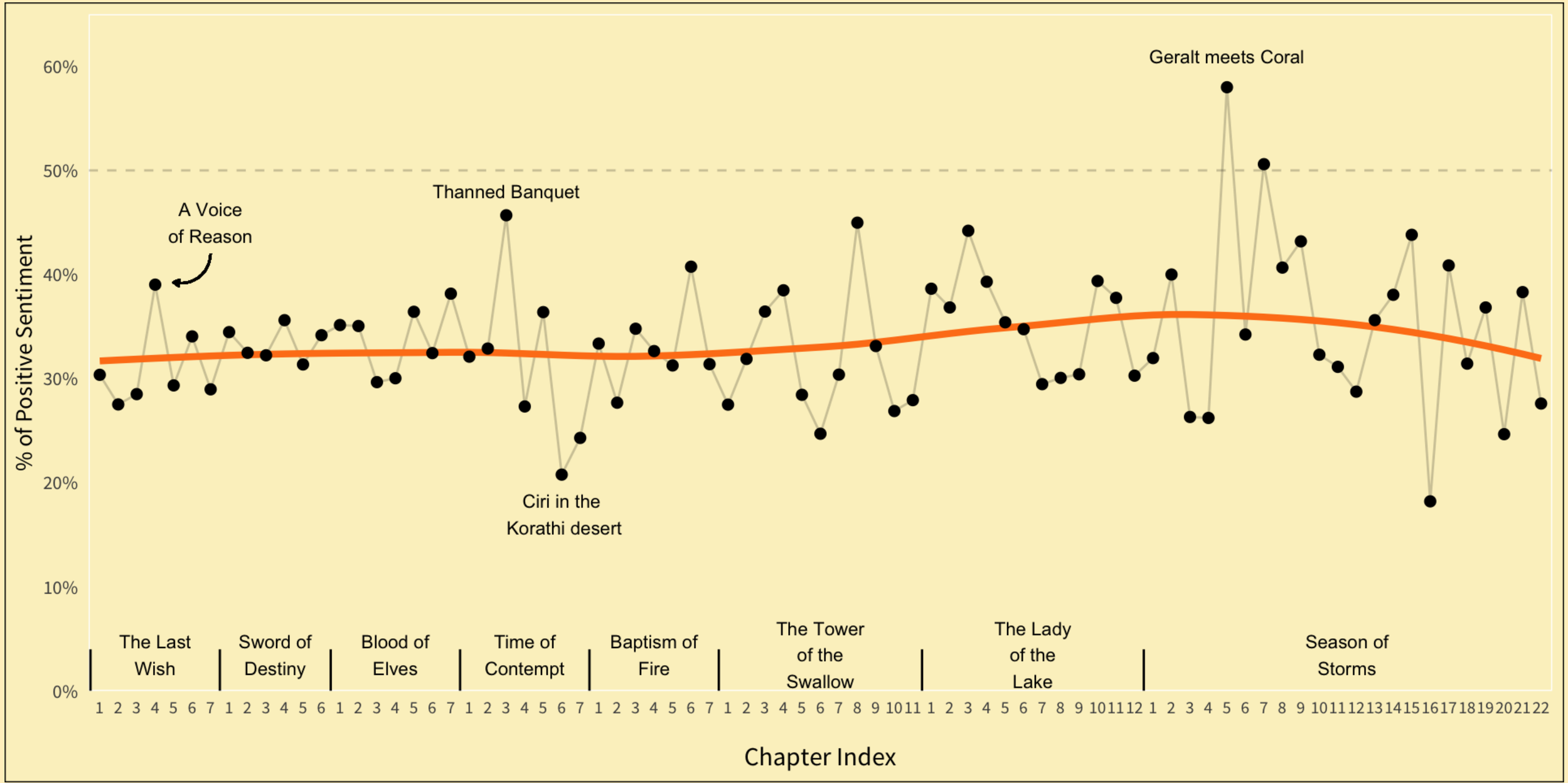
“Polokonfirmativní” analýza

Multimodální analýza

Online analýza

Analýza sentimentu

Analýza sentimentu Zaklínače (old school způsobem)



Topic modely jsou cool!

A to je vše, díky za pozornost!

https://github.com/alesvomacka/csr-scraping

ales@vomacka.io