Tematické modelování publikační činnosti České akademie věd a umění v letech 1890–1910

Diplomová práce

29. 1. 2024

Bc. Filip Kersch

Tematické modelování publikační činnosti České akademie věd a umění v letech 1890–1910

Cíl práce:

Na základě strojového analýzy tištěných publikací zjistit, jaká témata byla předmětem bádání České akademie císaře Františka Josefa pro slovesnost, vědy a umění v prvních dvaceti letech jejího fungování.

Obsah:

Česká akademie věd a umění

- Česká akademie císaře Františka Josefa pro vědy, slovesnost a umění vznikla 23. ledna 1890
- Organizována byla do 4 tříd, tato skutečnost měla dopad na publikační činnost:
  - společné publikace (*Věstník, Almanach),*
  - každá třída vydávala ***Rozpravy****,*
  - monografie,
  - podpora vydávání časopisů.

- **Témata (dosavadní stav poznání):**
  - I. třída: historie, právní problematika a filozofie,
  - II. třída: matematika, fyzika, chemii, biologie (bryologie), lékařství, technika,
  - III. filologie, čeština, česká literatura, lexikografie, dialektologie, orientální jazyky,…

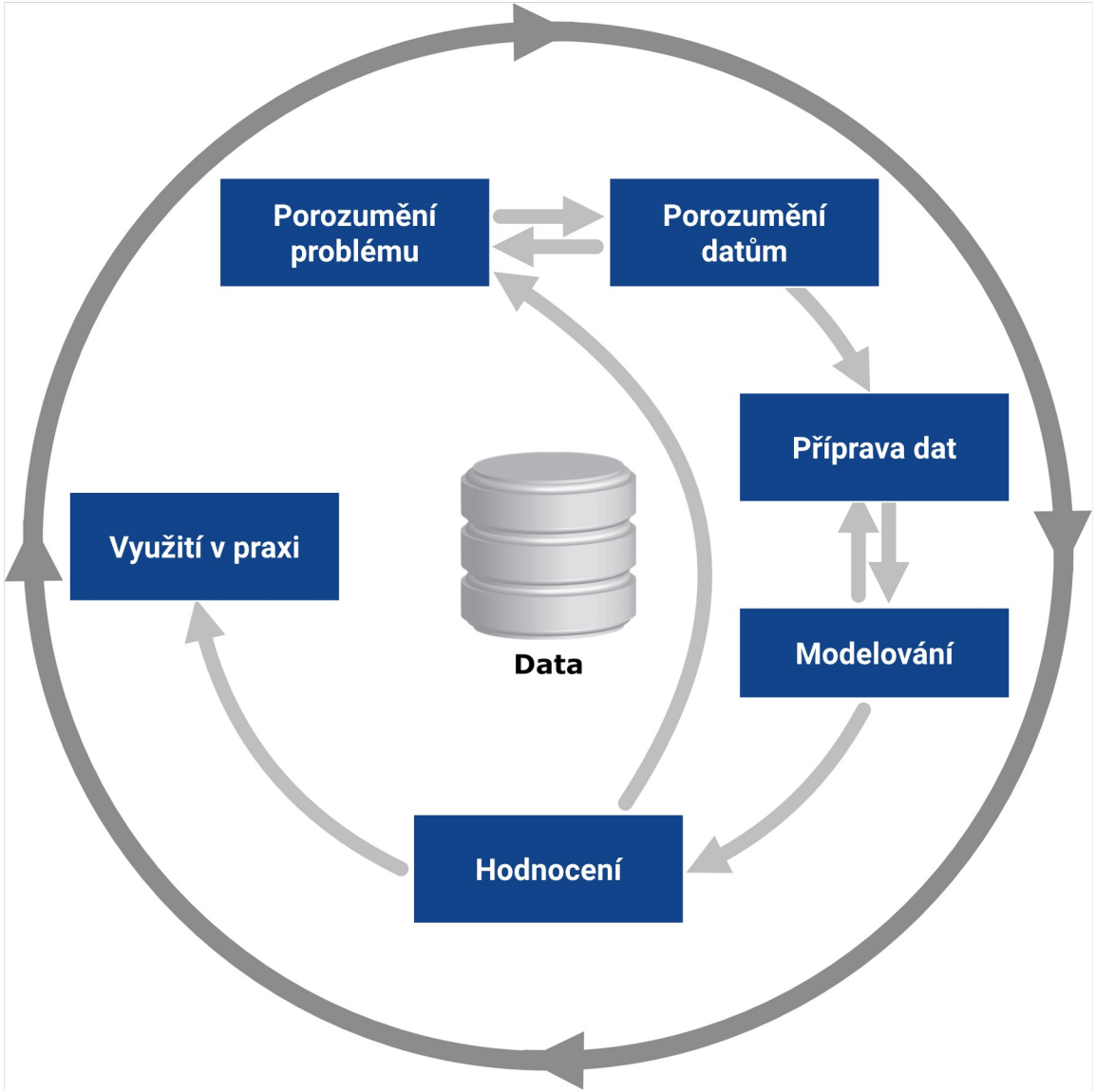
Digitalizace fondů
a makroanalýza digitálních dat

- Reformátování (digitalizace) *Rozprav* představuje nový zdroj pro analýzu.

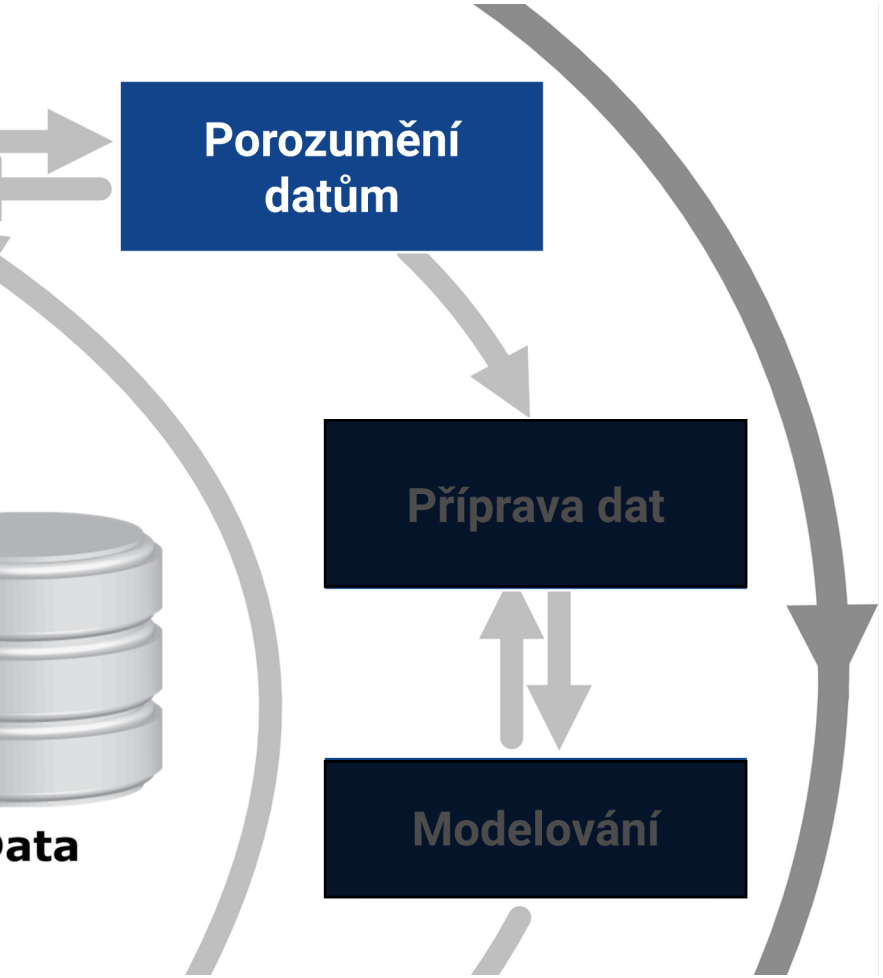
- Distant reading / makroanalýza:
  - Na texty v digitálních knihovnách lze nahlížet jako na data, ve kterých lze identifikovat vzorce a souvislosti.

- Tematické modelování
- **Latentní Dirichletova alokace**

Výzkumná část



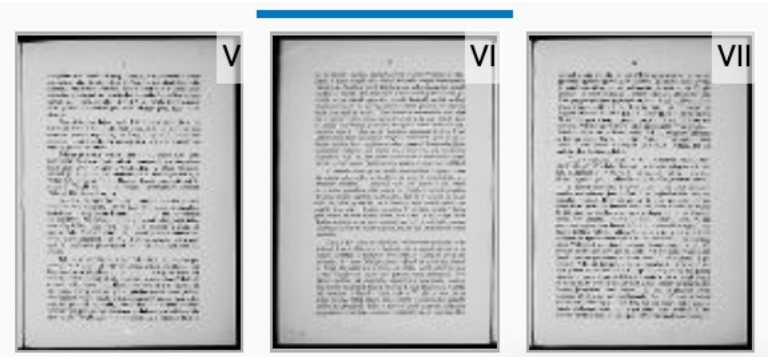
Porozumění datům



1892

Číslo 1

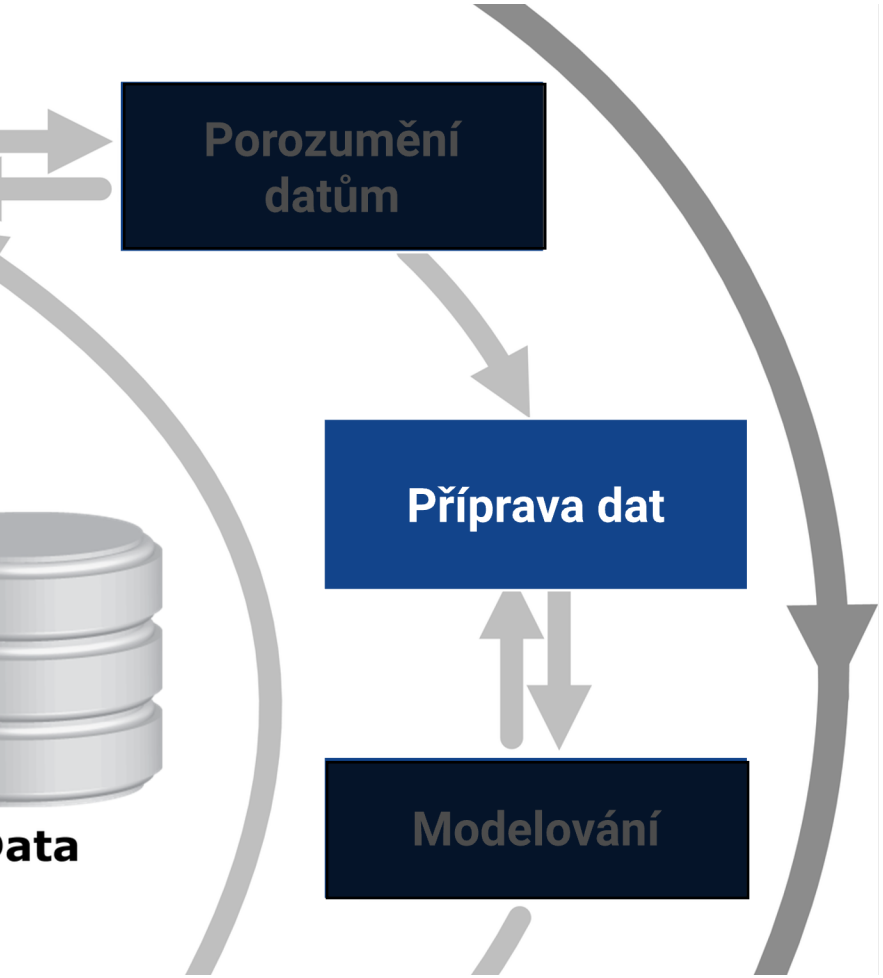
Digitální knihovna Akademie věd ČR

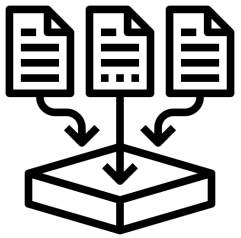


ocr | stránka | číslo periodika | ročník periodika | noviny | URL

PŘEDMLUVA.
Roku 1878 vydal jsem »Vergilstudien nebst einer Collati
V mnohých recensích vysloveno přání a očekávání, že pod
Toto pokračování podávám nyní, kdy zřízením České Akade
V práci této podány jsou k velikému počtu míst Vergilio
•'—1 Seznal jsem ovšem také tu i tam, ale celkem ve vel
l*

Příprava dat





.txt soubory:
887 svazků,
31 147 stran

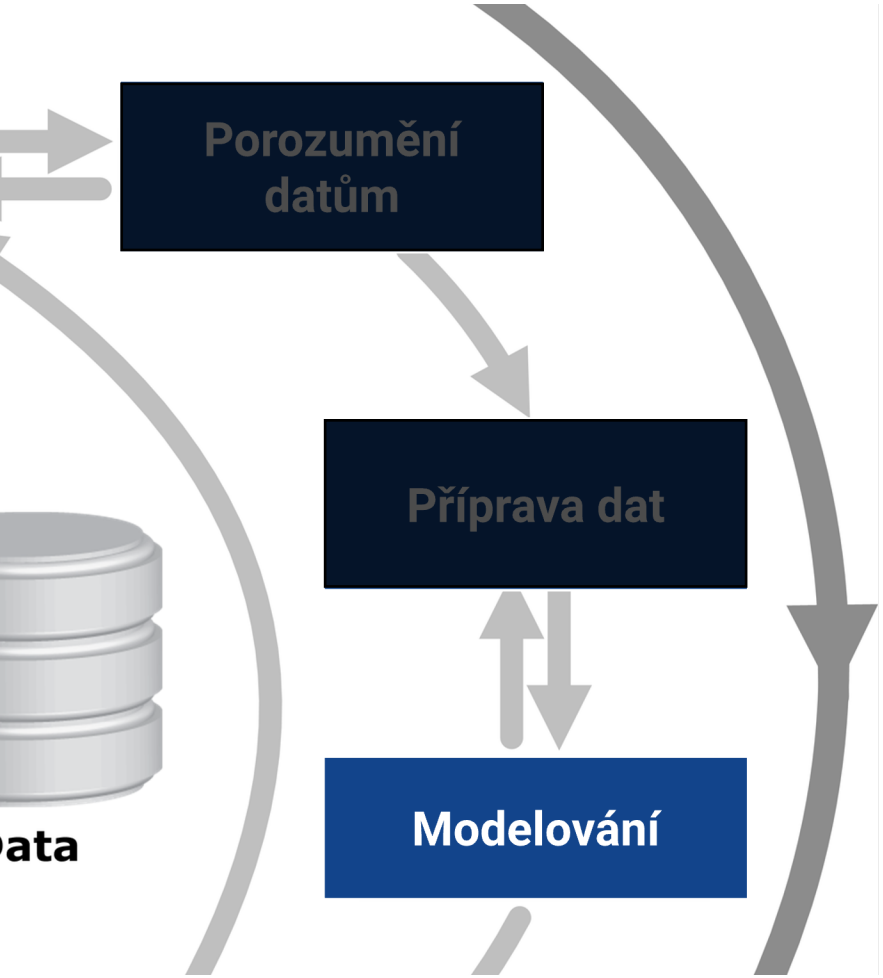
Pro přípravu dat i následné modelování bylo použito softwarové prostředí R.

Transformace, úpravy, obohacení:

- odstraněna slova obsahující dva nebo méně znaků,
- obohacení pomocí UDPipe 2 (tokenizace, lemmatizace a identifikaci slovního druhu),
- vyfiltrována jen podstatná jména.

Pro každý svazek *Rozprav* vznikl jeden .txt soubor obsahující pouze podstatná jména v základním tvaru

Modelování



- Vstupem pro model LDA je **matice dokument–slovo**

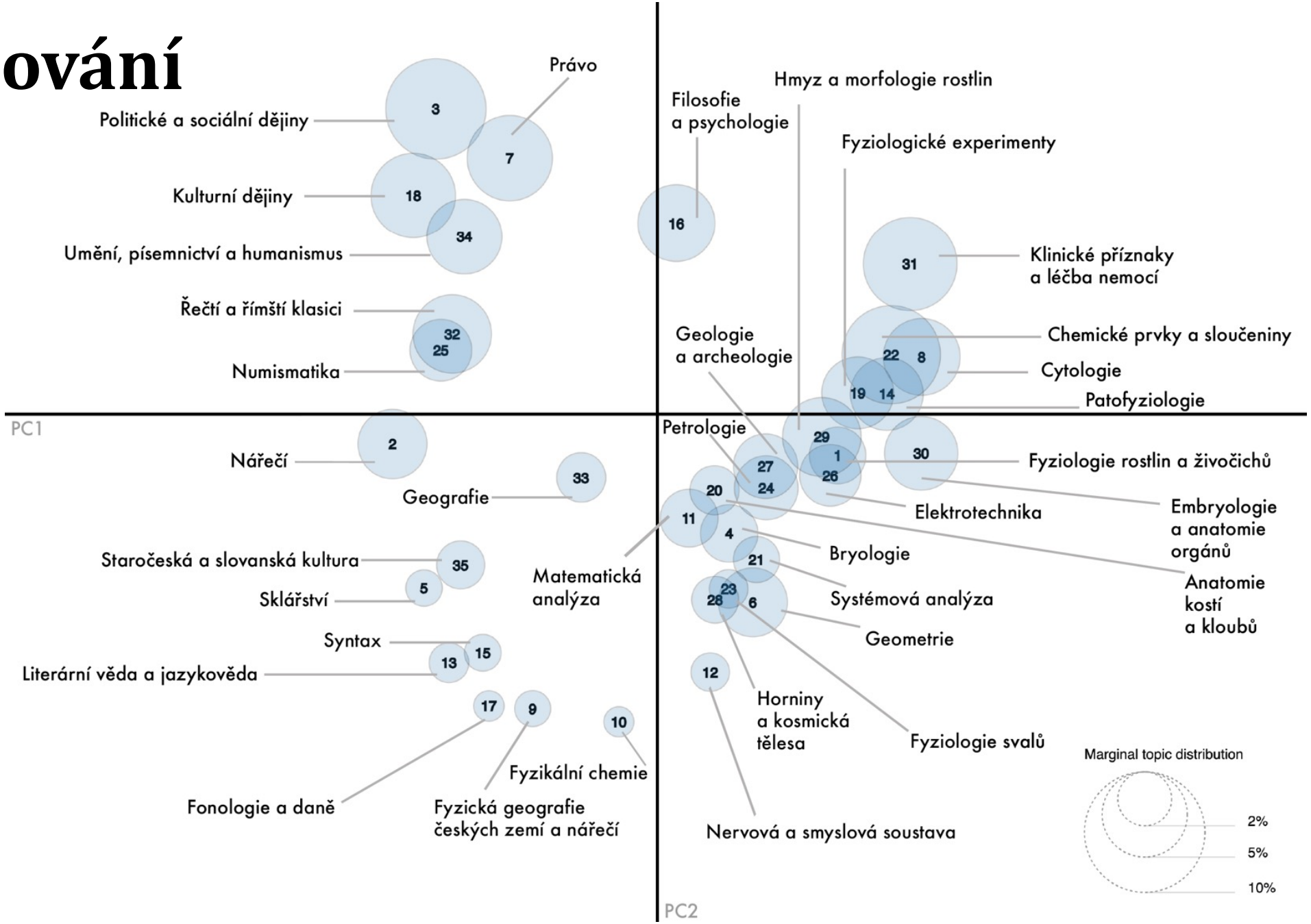
- K samotnému modelování témat byla využita knihovna `topicmodels`
  - Vyžaduje zvolit počet témat, která mají být v korpusu identifikována.
  - Na základě kvalifikovaného odhadu bylo zvoleno 35 témat.
  - **Cílem modelování je identifikovat témata s alespoň o úroveň jemnější granularitou, než doposud.**

Modelování

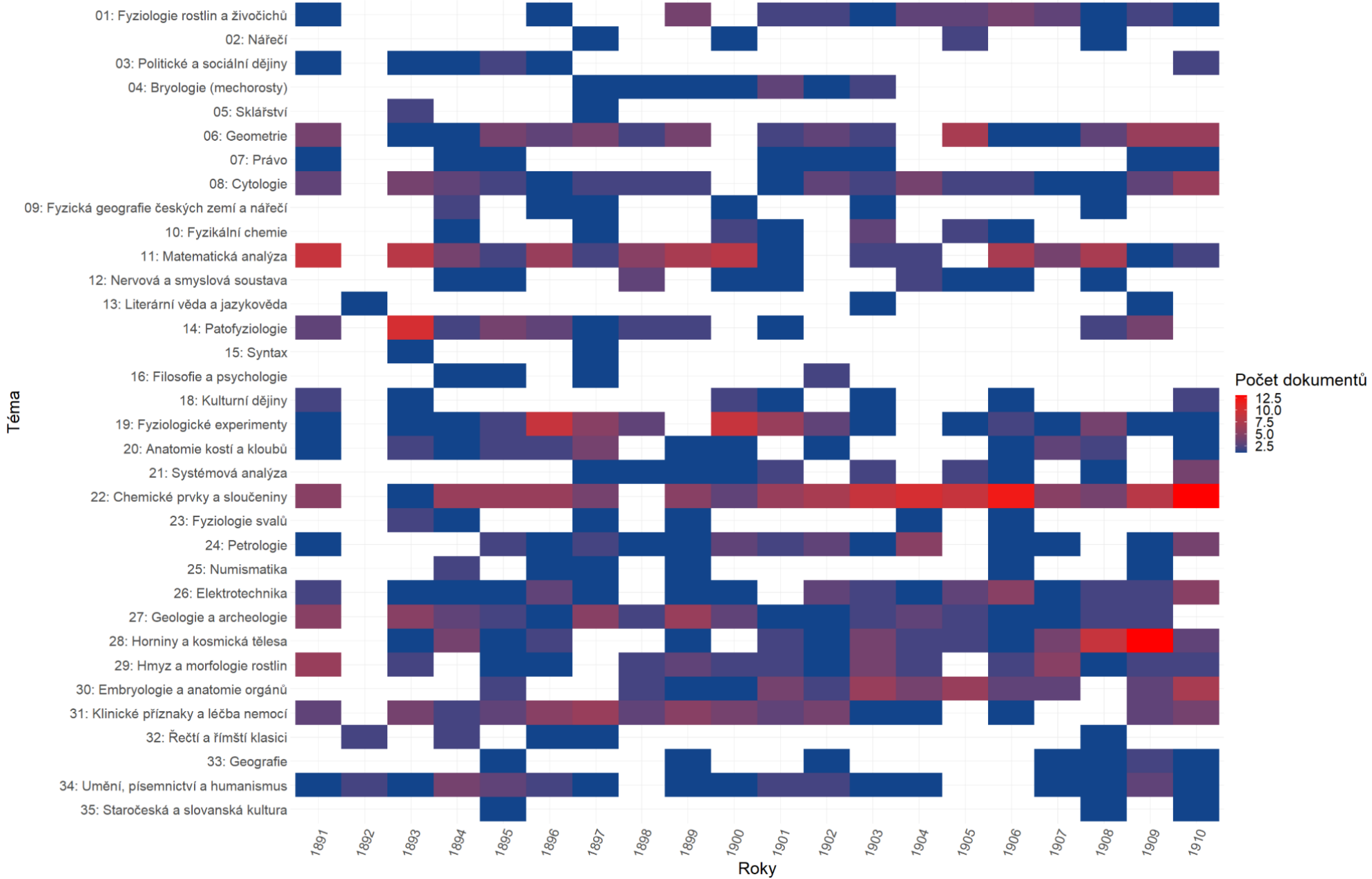
- Výstupem modelu jsou dvě matice:
  - **slovo–téma**, ve které je u každého slova z korpusu uvedeno,
    s jakou pravděpodobností „patří“ k jakému tématu.
  - **dokument–téma** popisující pravděpodobnost příslušnosti dokumentu
    ke konkrétnímu tématu.

- **Model rozpoznaná témata v korpusu označuje pouze číslem**,
  pro pojmenování byly využity tři techniky:
  - nejfrekventovanější slova, která se v každém tématu vyskytují (17 případů),
  - analýza začátků dokumentů patřících k danému tématu s nejvyšší pravděpodobností,
  - analýza zdrojových digitalizátů.

Modelování


Právo
Politické a sociální dějiny
3
7
Kulturní dějiny
18
34
Umění, písemnictví a humanismus
Řečtí a římští klasici
32
25
Numismatika
Hmyz a morfologie rostlin
Filosofie
a psychologie
Fyziologické experimenty
16
31
Klinické příznaky
a léčba nemocí
Geologie
a archeologie
Chemické prvky a sloučeniny
22
8
Cytologie
19
14
Patofyziologie
PC1
Petrologie
29
1
30
Fyziologie rostlin a živočichů
27
24
26
Elektrotechnika
Embryologie
a anatomie
orgánů
2
Nářečí
33
Geografie
20
11
4
Bryologie
Staročeská a slovanská kultura
35
Matematická
analýza
21
Anatomie
kostí
a kloubů
5
Sklářství
23
28
6
Systémová analýza
Geometrie
Syntax
15
13
Literární věda a jazykověda
12
Horniny
a kosmická
tělesa
17
9
10
Fyzikální chemie
Fyziologie svalů
Fonologie a daně
Fyzická geografie
českých zemí a nářečí
Nervová a smyslová soustava
Marginal topic distribution
2%
5%
10%
PC2

Modelování


01: Fyziologie rostlin a živočichů
02: Nářečí
03: Politické a sociální dějiny
04: Bryologie (mechorosty)
05: Sklářství
06: Geometrie
07: Právo
08: Cytologie
09: Fyzická geografie českých zemí a nářečí
10: Fyzikální chemie
11: Matematická analýza
12: Nervová a smyslová soustava
13: Literární věda a jazykověda
14: Patofyziologie
15: Syntax
16: Filosofie a psychologie
18: Kulturní dějiny
19: Fyziologické experimenty
20: Anatomie kostí a kloubů
21: Systémová analýza
22: Chemické prvky a sloučeniny
23: Fyziologie svalů
24: Petrologie
25: Numismatika
26: Elektrotechnika
27: Geologie a archeologie
28: Horniny a kosmická tělesa
29: Hmyz a morfologie rostlin
30: Embryologie a anatomie orgánů
31: Klinické příznaky a léčba nemocí
32: Řečtí a římští klasici
33: Geografie
34: Umění, písemnictví a humanismus
35: Staročeská a slovanská kultura
Téma
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
Roky
Počet dokumentů
12.5
10.0
7.5
5.0
2.5

Výsledky

- Provedená analýza podává přehled **konkrétních témat, kterým se Česká akademie ve sledovaném období věnovala**,
  - jejich sémantické podobnosti,
  - počtu svazků zařazených do každého z témat
  - a proměn počtu vydaných svazků v jednotlivých letech.

- Rozpoznaná témata byla namapována na kategorie vycházející z vědních sekcí AV ČR a také vědních oborů a oborových skupin FORD.
- Práce **představuje podrobný pohled na publikační činnost České akademie jako celek** a umožňuje rozšířit dosavadní stav poznání oblastí zájmů této instituce.

Závěr

- Součástí diplomové práce jsou 3 přílohy:
  - **Příloha č. 1:** Seznam publikací vydaných Českou akademií věd císaře Františka Josefa pro vědy, slovesnost a umění z období let 1890–1918.
  - **Příloha č. 2:** Postup stažení a přípravy dat a modelování témat
    - Data a scripty byly zveřejněny také na platformě GitHub: https://github.com/kerschfilip/tematicke_modelovani_cavu
    - Veškerá data jsou dostupná v repozitáři Zenodo: https://doi.org/10.5281/zenodo.10395970
  - **Příloha č. 3**: Seznam odkazů na digitalizované svazky časopisu Rozpravy rozřazený podle rozpoznaných témat

Tematické modelování publikační činnosti České akademie věd a umění v letech 1890–1910

Diplomová práce

29. 1. 2024

Bc. Filip Kersch