Top2Vec as a tool for topic identification in fiction. A study on contemporary Czech prose.

8. 02. 2024 Charlotte Panušková

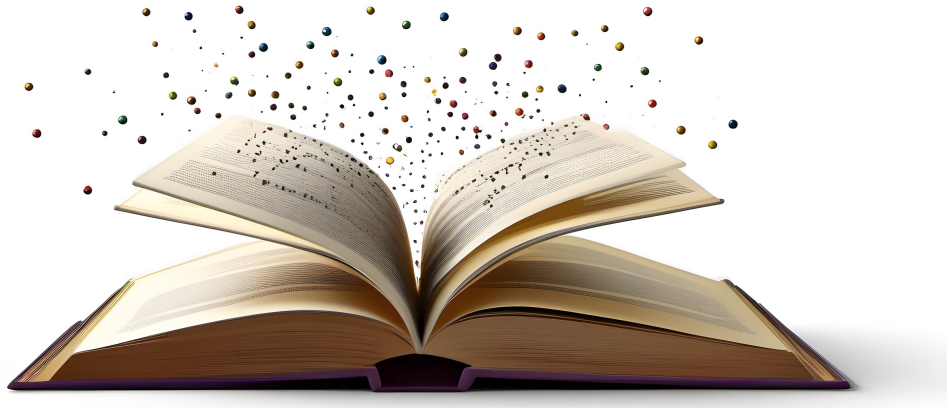
Ústav pro českou literaturu AV ČR
Institute of Czech Literature of the CAS

Outline

1. Introduction
2. Background
3. Research Methodology
4. Topic Modelling Techniques
5. Results and Observations
6. Conclusion
7. Q&A

Research question

- Trends in contemporary Czech literature -> topics?
- How does topic modelling relate to traditional literary studies?



Trends in Czech literature

1990s - Two streams emerged

- fantasy, imaginative prose associated with the metropolis (mostly Prague)
- authenticity, connected by the author's "self"

2000s - Works set in specific times and places

- historical traumas (holocaust, expulsion of Germans ...)
- the atmosphere of life in (post)communist society

2010s – Continuation of the trend from the previous decade

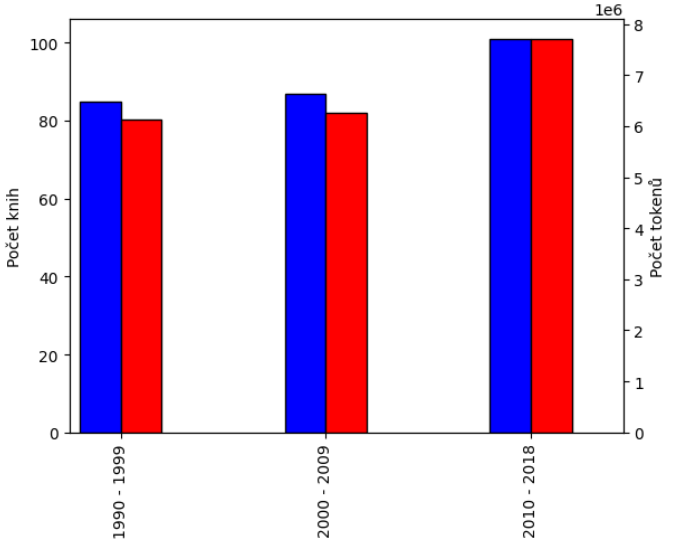
- Relationship crisis
- (Auto)biographies

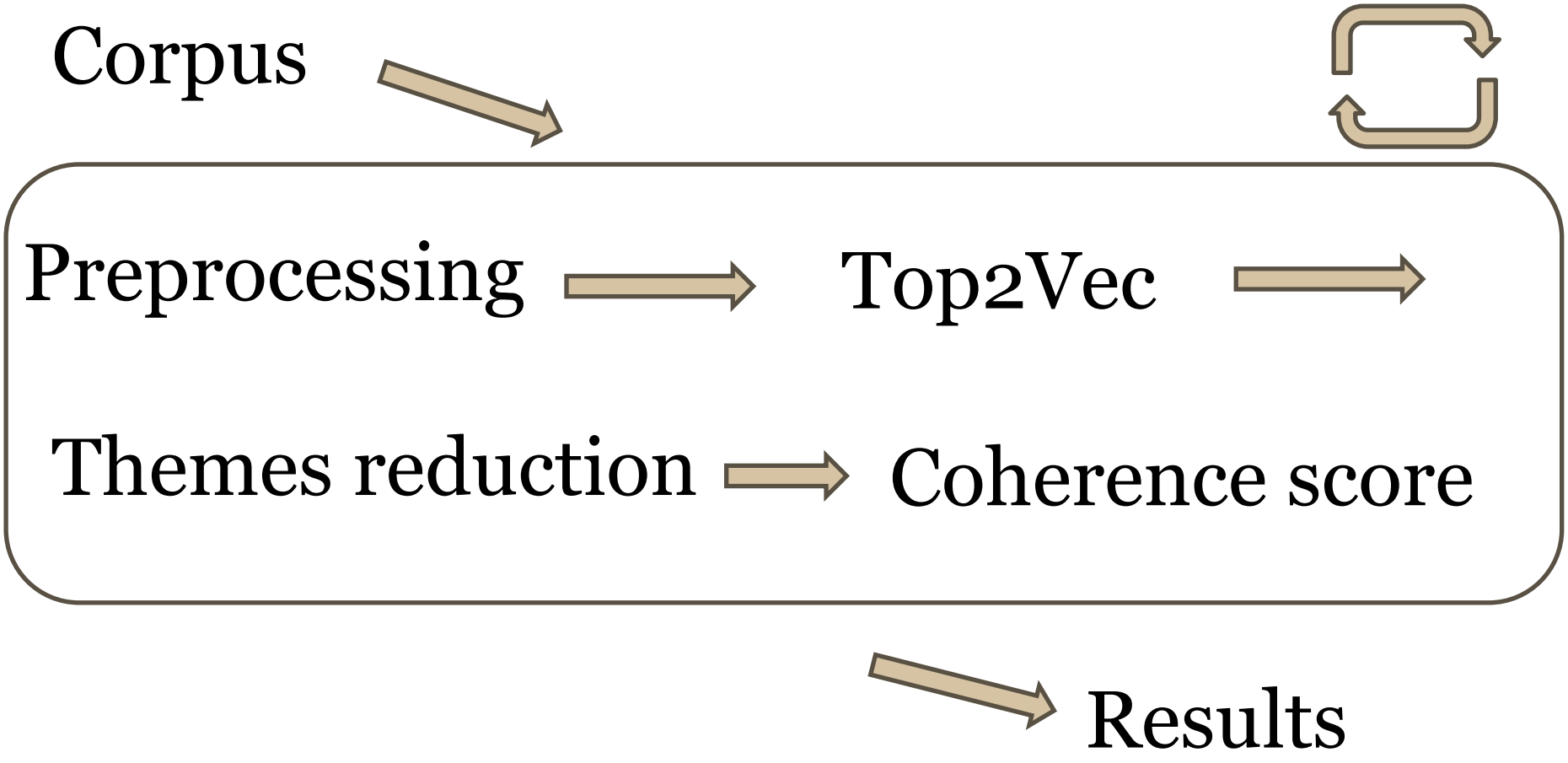
Topic Modelling in DLS

- Jockers, Matthew L., and David Mimno. ‘Significant Themes in 19th-Century Literature’, 2013.
- Schöch, Christof. ‘Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama’, 2021.
- Van Zundert, Joris J et. al. ‘What Do We Talk About When We Talk About Topic?’, 2022.

Corpus

- Publicly available corpus from the Czech National Corpus, SYN v9
- Only fiction -> prose
- Not representative, however diverse
- documents split into sections and mixed
- 273 books
- Divided into 3 decades based on the date of first publication



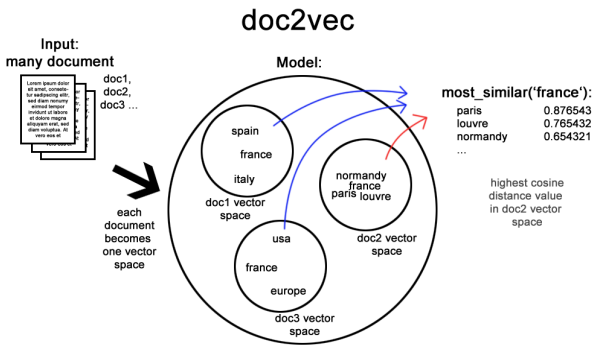

Corpus
Preprocessing
Top2Vec
Themes reduction
Coherence score
Results

Preprocessing

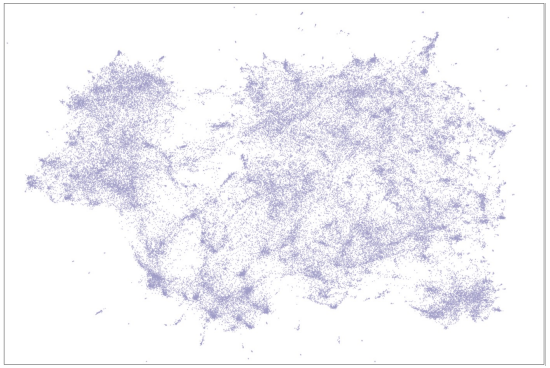
- The corpus is already lemmatized.
- Removal of Named Entity Recognition (NER) - NameTag
- Removal of stopwords.
- Splitting documents into parts.

Top2Vec

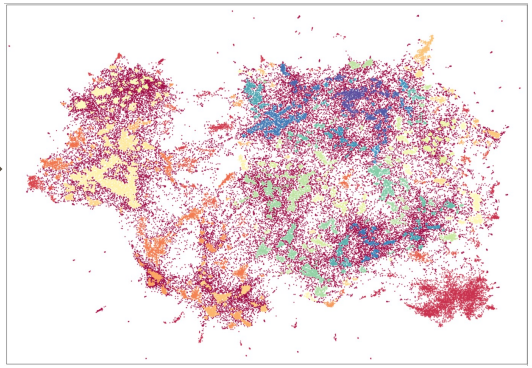
doc2vec



UMAP



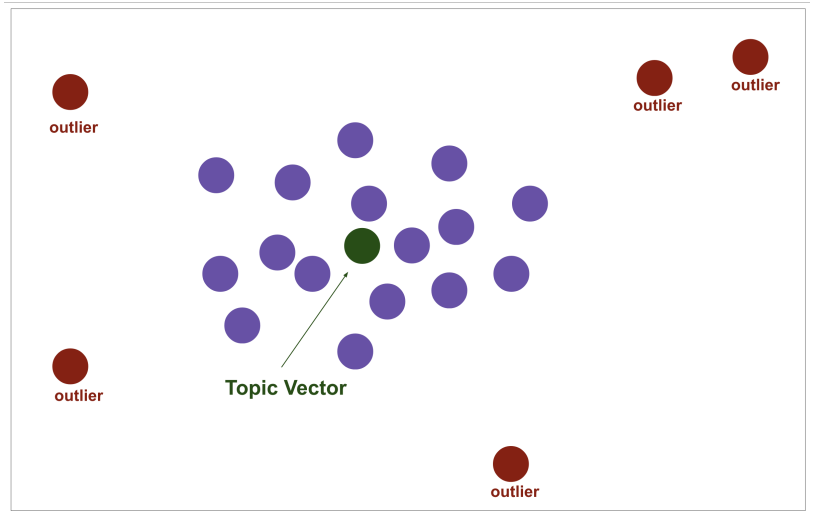
HDBSCAN

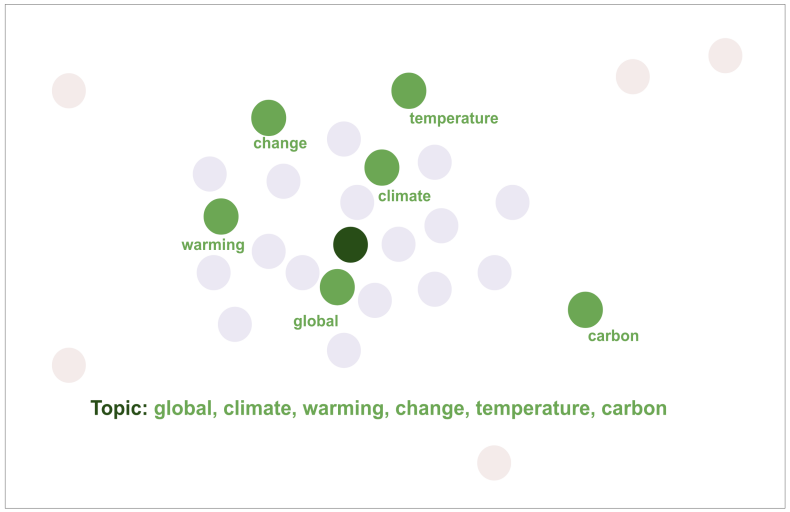


Pham, Hieu Huu Quang. (2017). An Empirical Approach to Sentiment Analysis with Doc2Vec. Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/190914.

Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

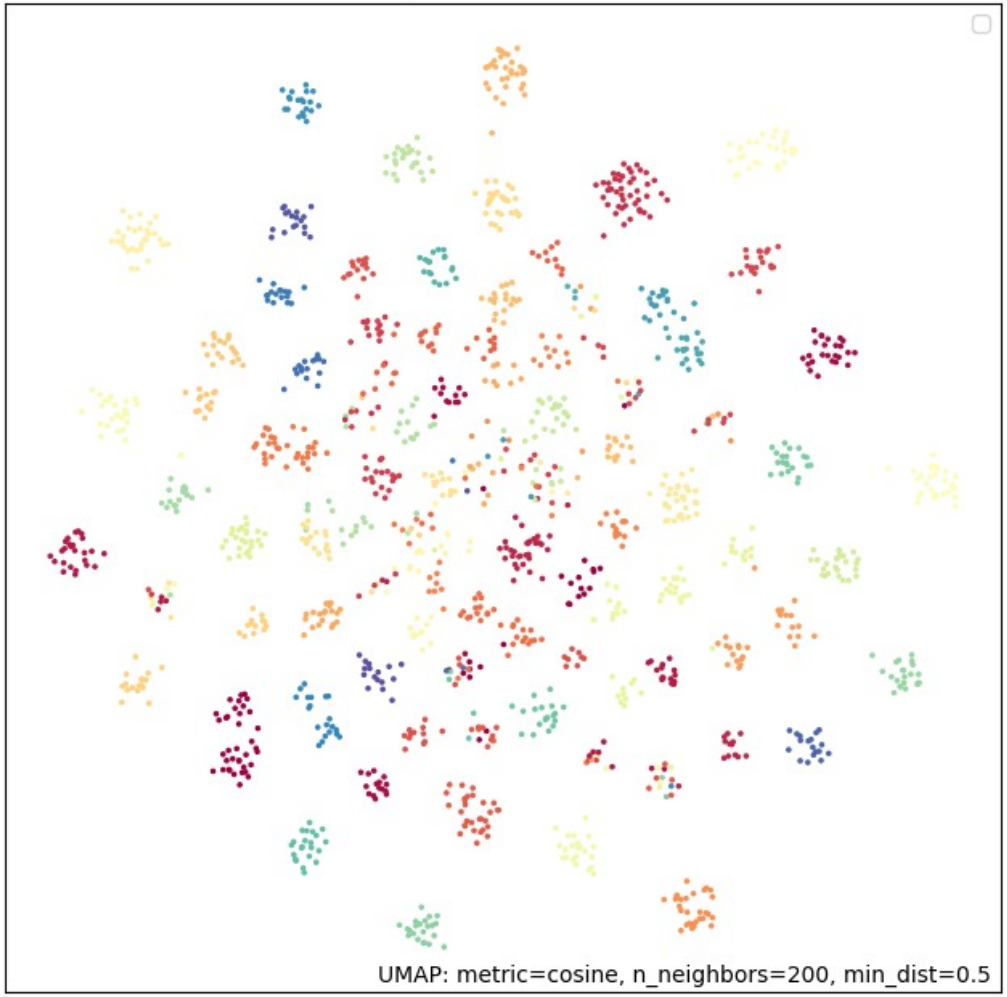
Top2Vec





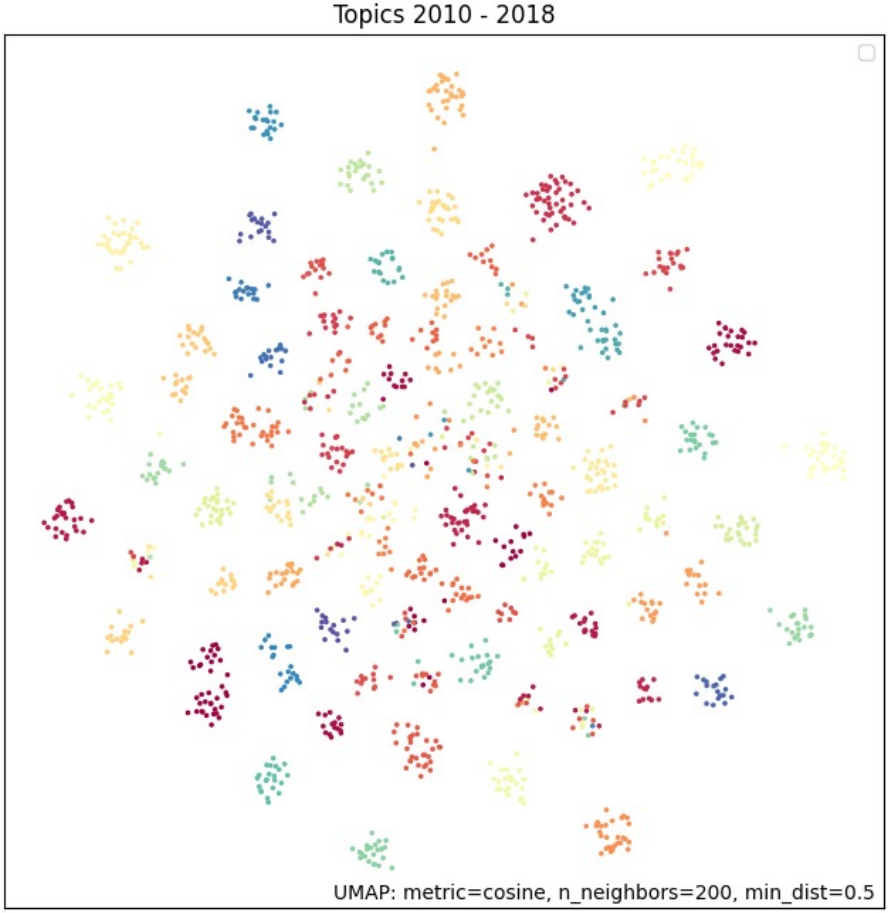
Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

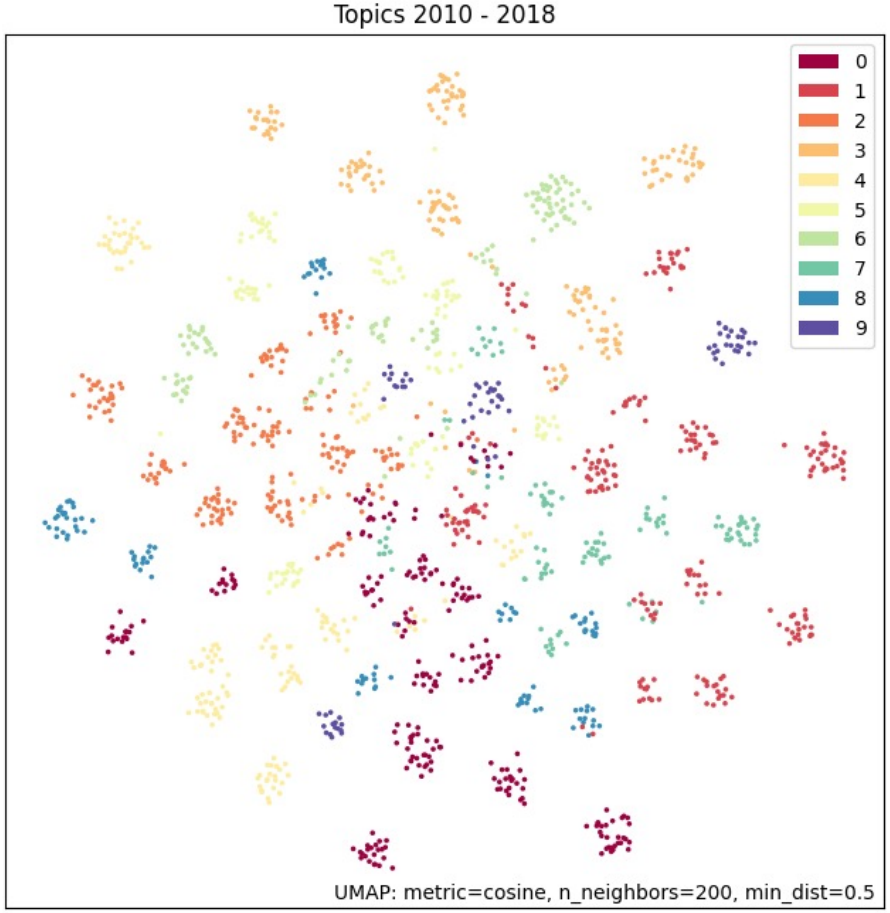
Topics 2010 - 2018



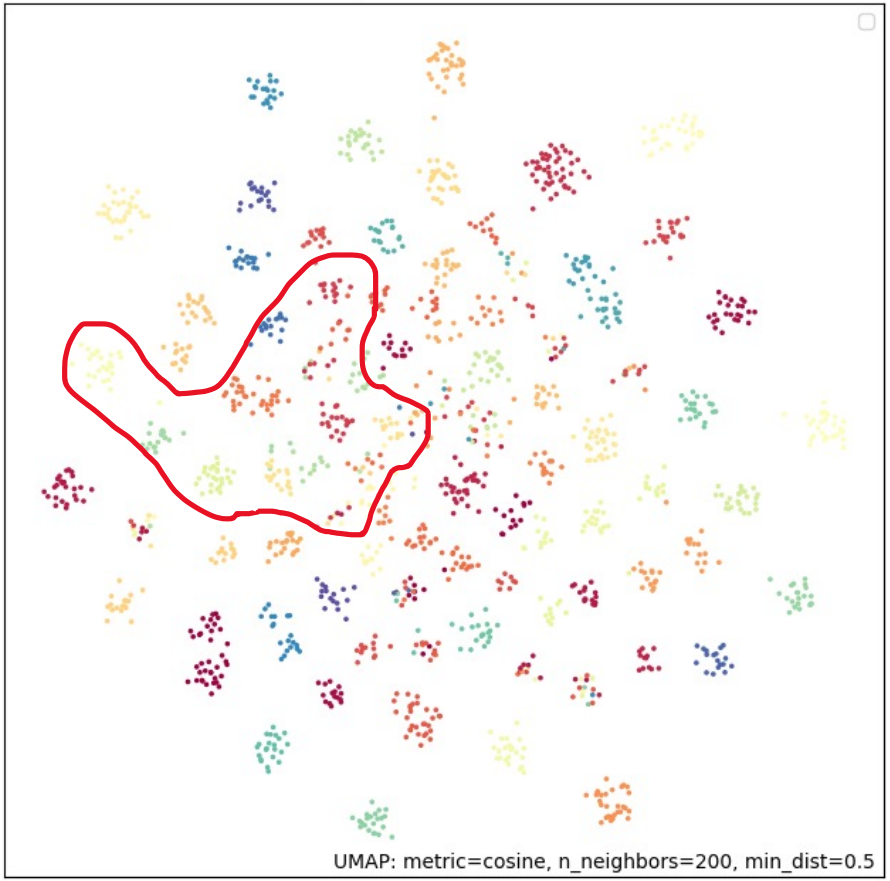
Coherence score

Tokens	Topics	C_v
1000	10	0,447
1000	20	0,465
1000	30	0,471
2000	**10**	**0,479**
2000	20	0,470
2000	30	0,468

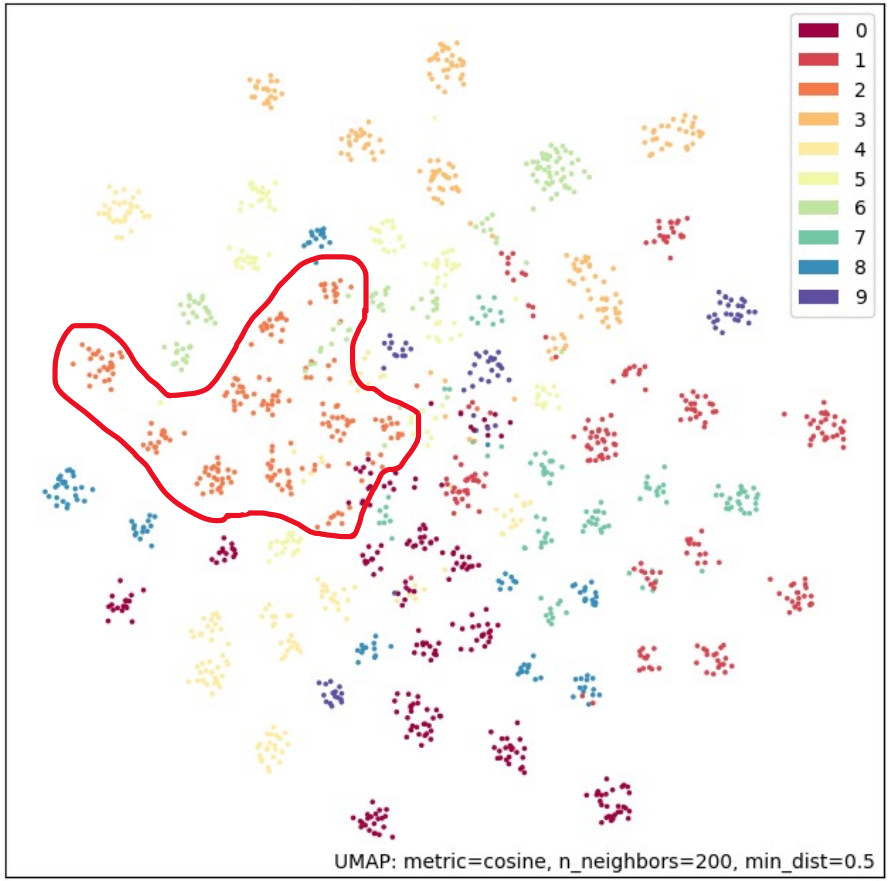

Topics 2010 - 2018
UMAP: metric=cosine, n_neighbors=200, min_dist=0.5


Topics 2010 - 2018
0
1
2
3
4
5
6
7
8
9
UMAP: metric=cosine, n_neighbors=200, min_dist=0.5

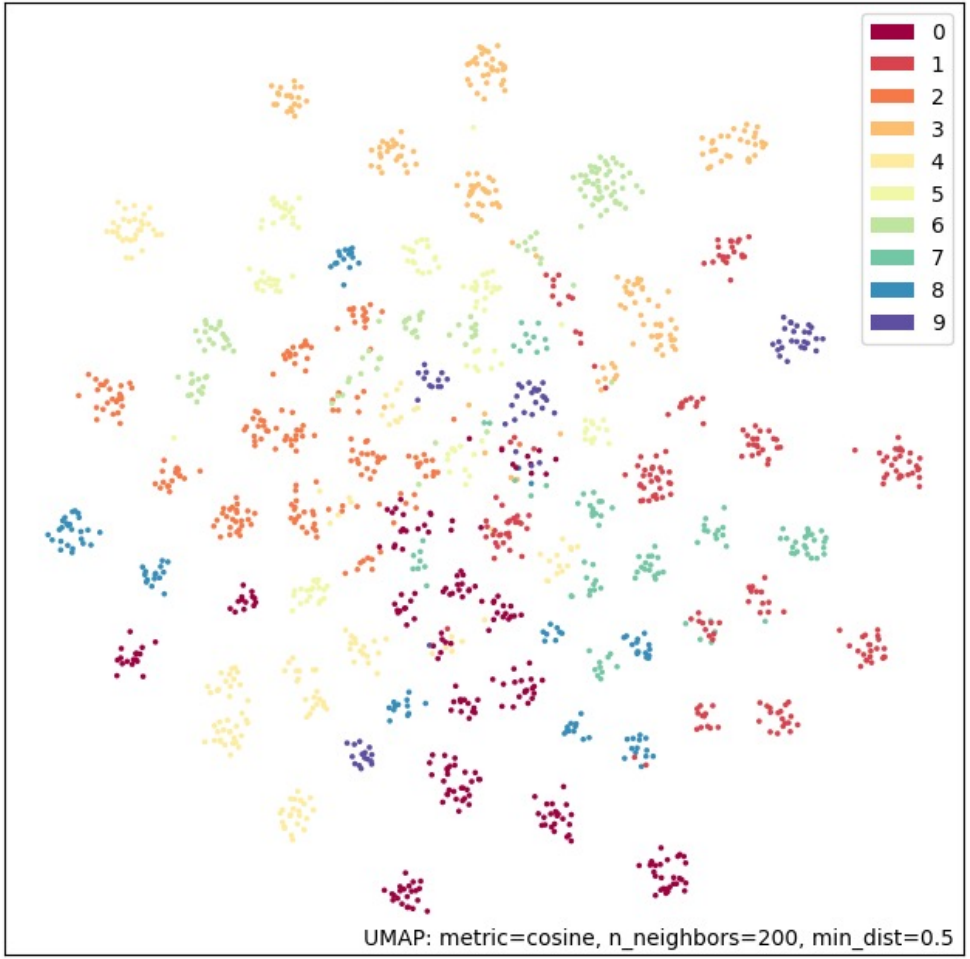
Topics 2010 - 2018



Topics 2010 - 2018

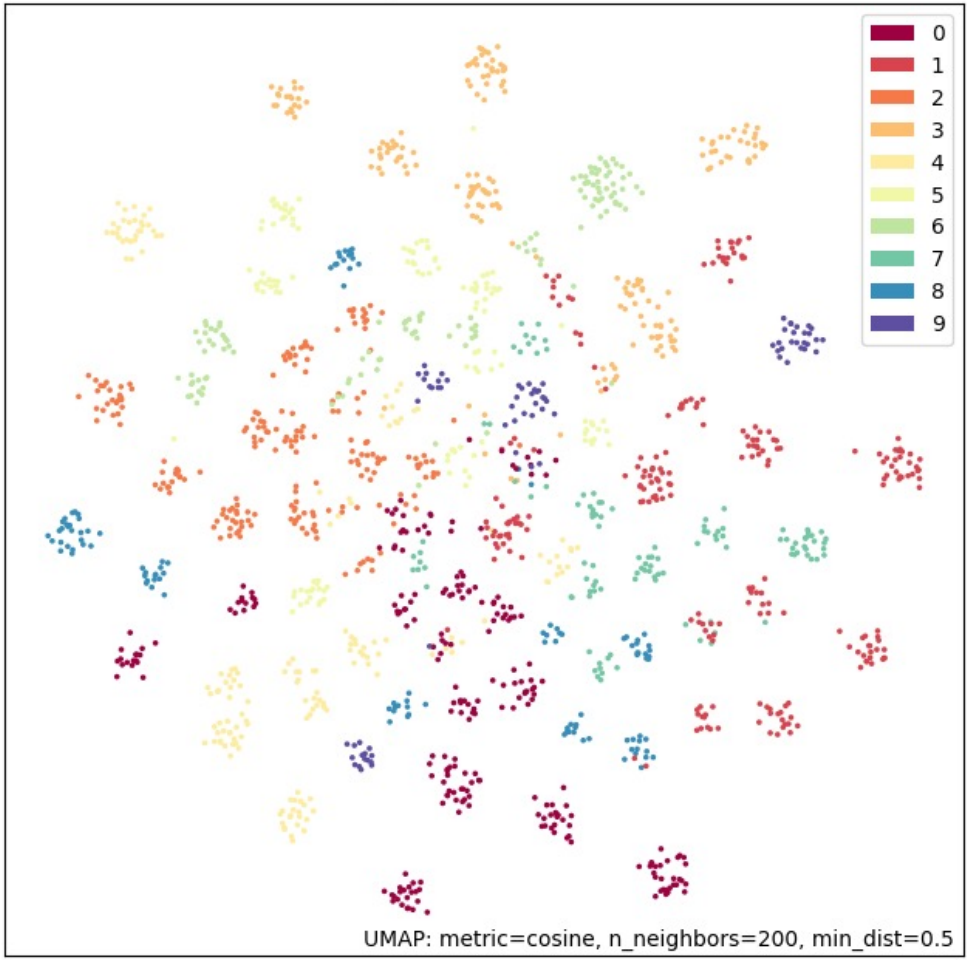


Topics 2010 - 2018



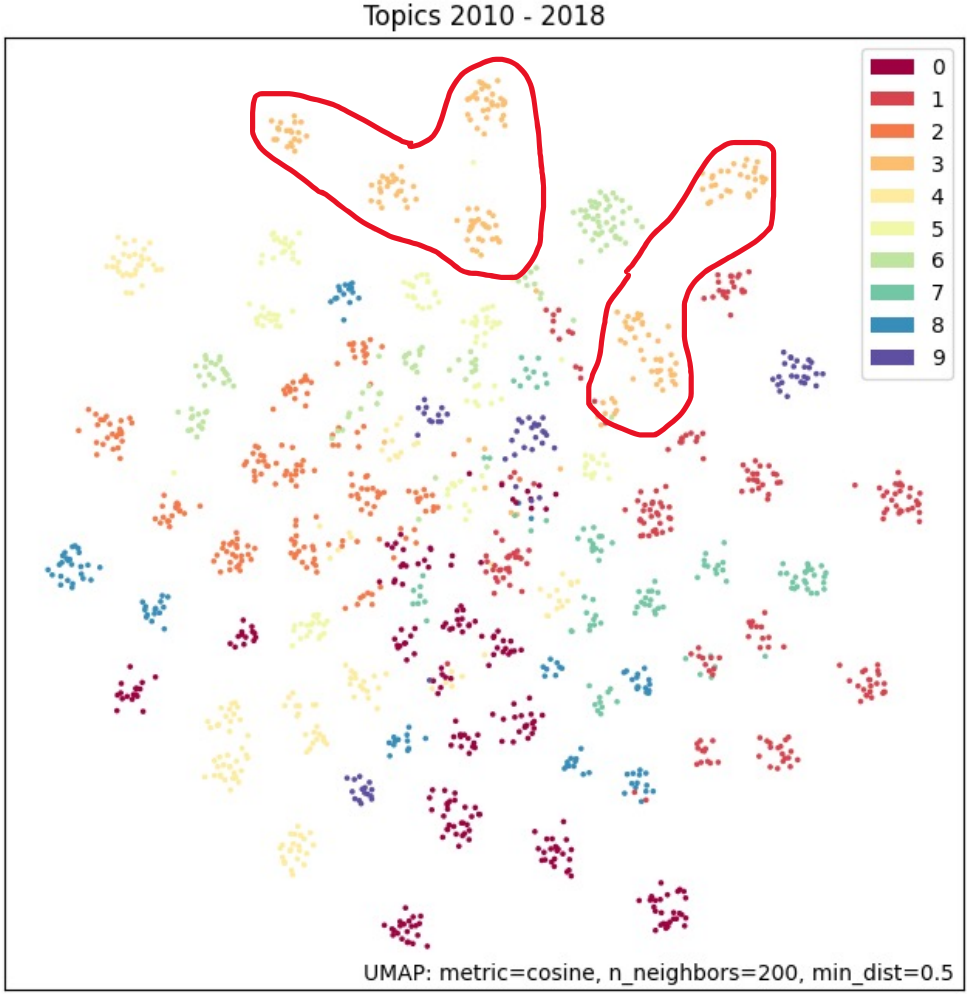
3
communist
mom, mother
communist
party-related
comrade
military service
resistance
socialist
Czechoslovakian
regime

Topics 2010 - 2018



3
komunista
maminka
komunisticky
stranicky
soudruh
vojna
odbojar
socialisticky
ceskoslovensky
rezim

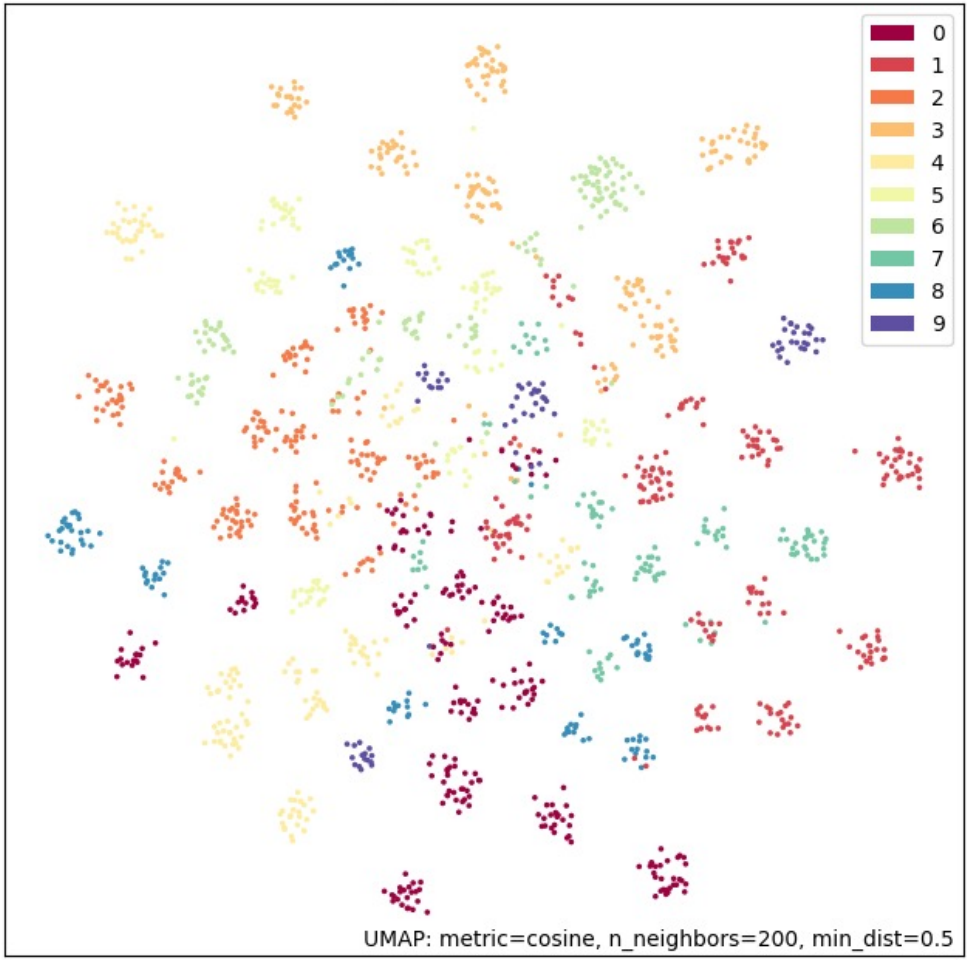
3
communist
mom, mother
communist
party-related
comrade
military service
resistance
socialist
Czechoslovakian
regime



3
komunista
maminka
komunisticky
stranicky
soudruh
vojna
odbojar
socialisticky
ceskoslovensky
rezim

4
dad
uncle
village
cottage
this year
village square
holiday
bag
Lhotka
bus

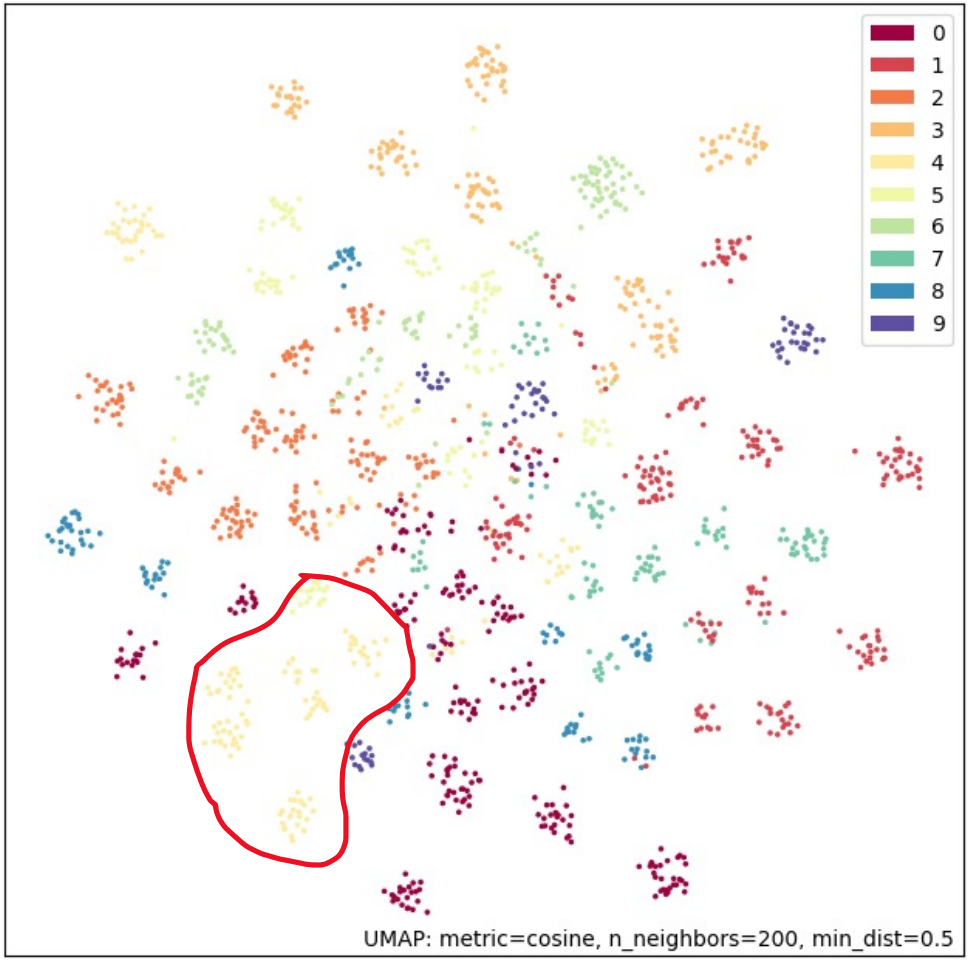
Topics 2010 - 2018



4
tata
strejda
ves
chalupa
letos
naves
prazdniny
taska
lhotka
autobus

4
dad
uncle
village
cottage
this year
village square
holiday
bag
Lhotka
bus

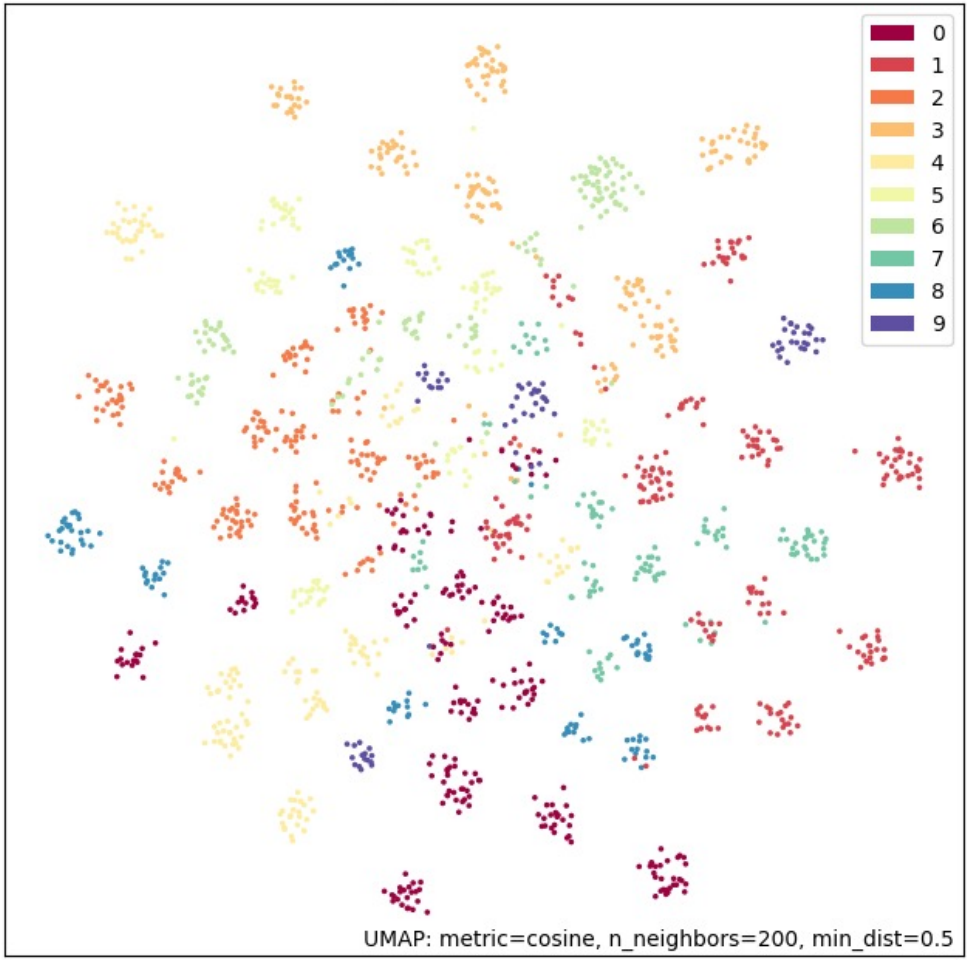
Topics 2010 - 2018



4
tata
strejda
ves
chalupa
letos
naves
prazdniny
taska
lhotka
autobus

7
heretic
duke
divine
church
archbishop
master
saint
tavern
teaching
parish priest

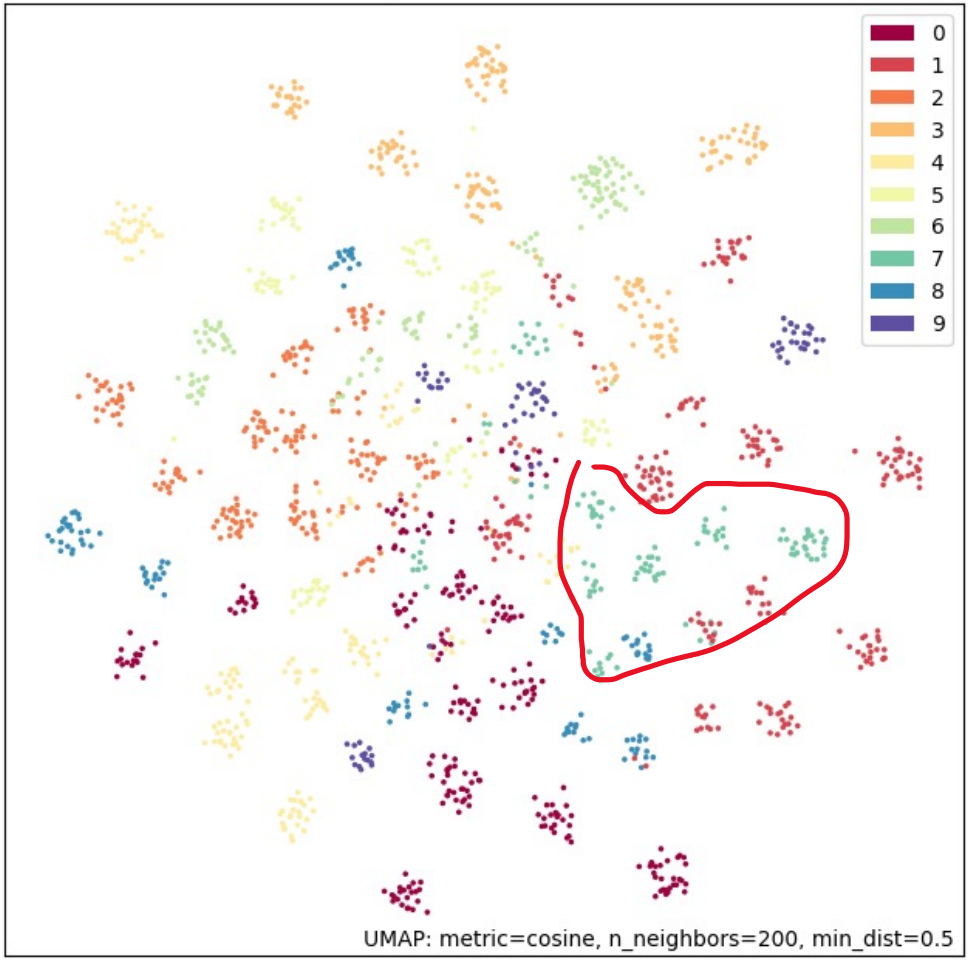
Topics 2010 - 2018



7
kacir
knez
bozi
cirkev
arcibiskup
mistr
svaty
krcma
uceni
farar

7
heretic
duke
divine
church
archbishop
master
saint
tavern
teaching
parish priest

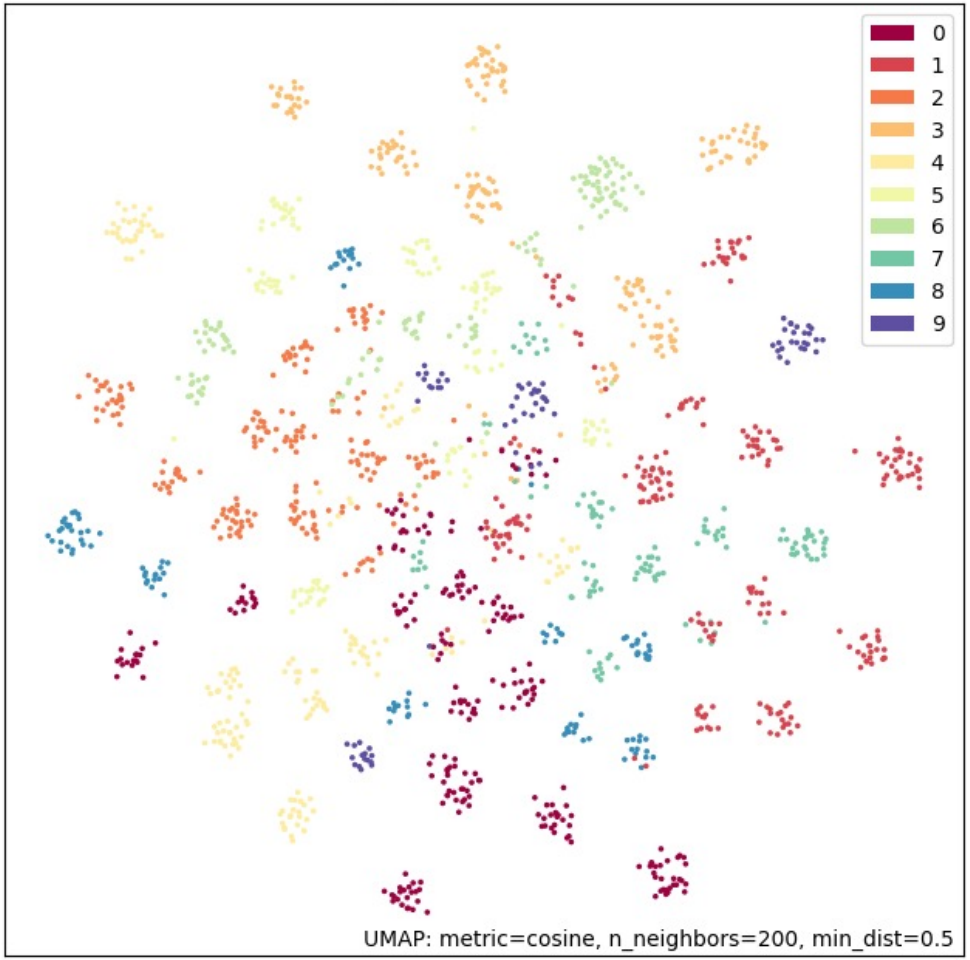
Topics 2010 - 2018



7
kacir
knez
bozi
cirkev
arcibiskup
mistr
svaty
krcma
uceni
farar

9
goddess
hillside
swallow
body
policeman's
birdie
postman
research
yoga
burrow

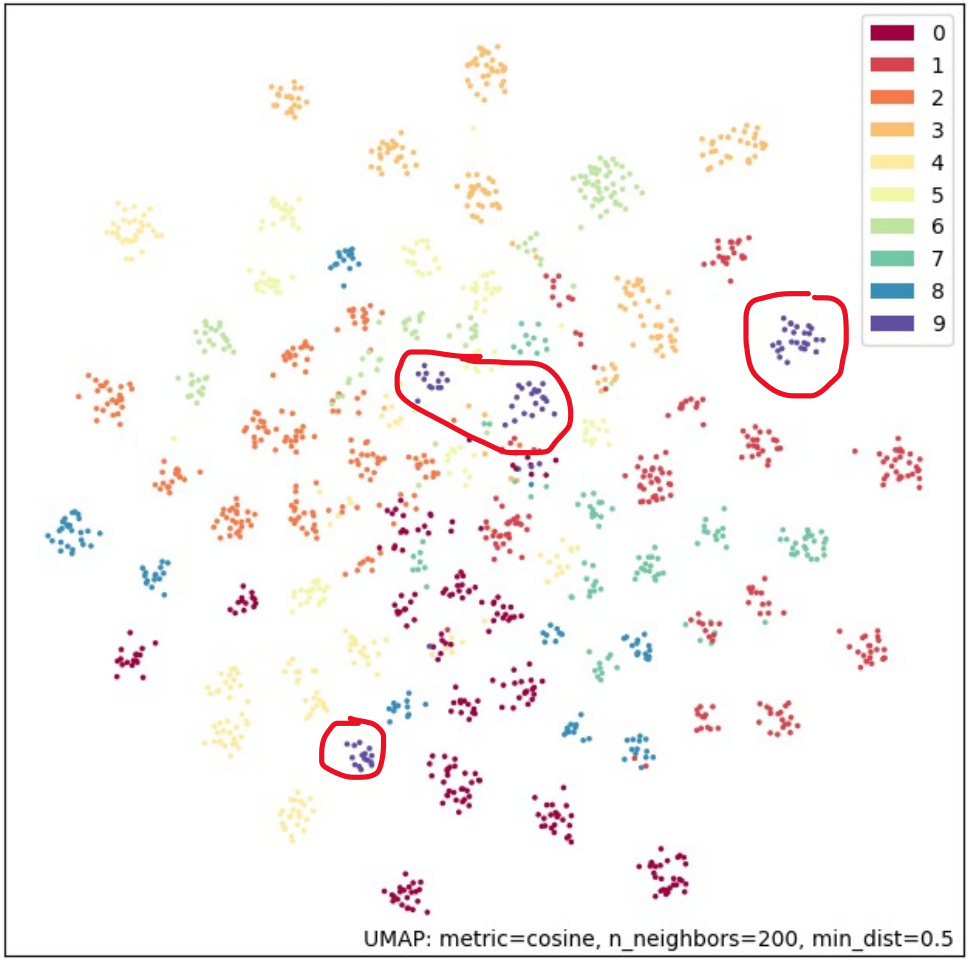
Topics 2010 - 2018



9
bohyne
kopanice
vlastovka
telo
policistuv
ptacek
postak
vyzkum
joga
doupe

9
goddess
hillside
swallow
body
policeman's
birdie
postman
research
yoga
burrow

Topics 2010 - 2018

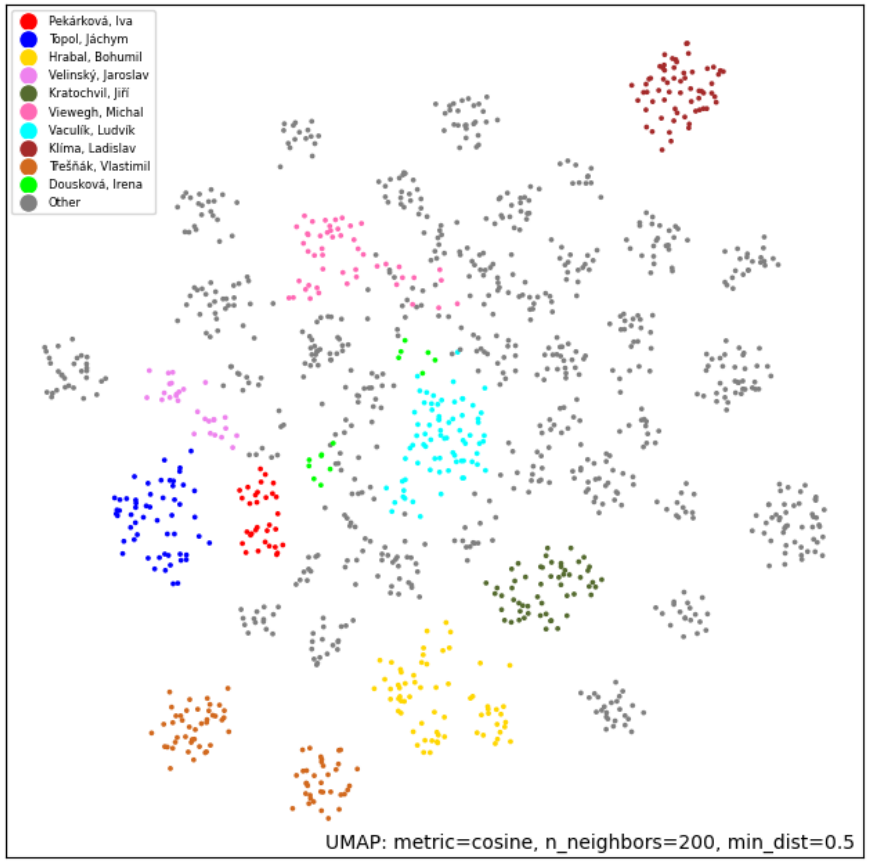


9
bohyne
kopanice
vlastovka
telo
policistuv
ptacek
postak
vyzkum
joga
doupe

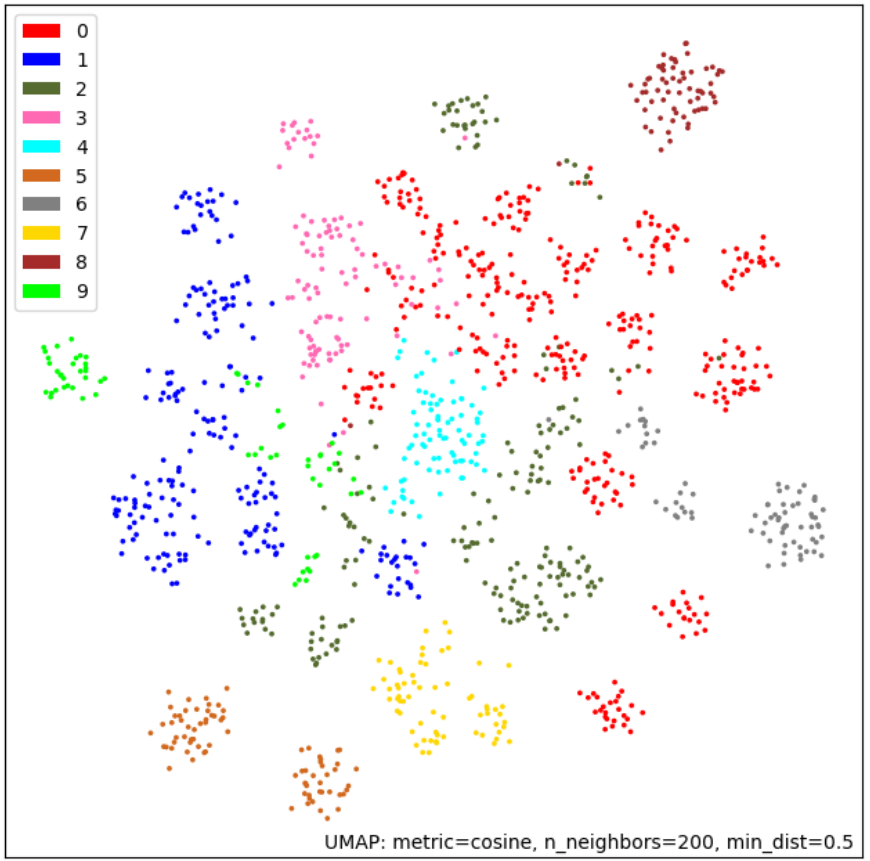
Observation

- Themes are fairly coherent
- Words are not always representative; it is necessary to look at specific works.
- Problem of dividing books into parts – the longer the book, the greater the weight.
- Authors stay within one topic cluster

Authors with Multiple Works in the Corpus 1990 - 1999



Topic clusters 1990 - 1999



Results

- What is a ‘topic' according to Top2Vec?
  - Motif, Theme
  - Genre
  - Author’s style
- Literary system?
- The words of the topics corresponded to literary trends only if the trend was clearly defined (eg. Communism)

Future work

- Representative corpus
- More robust model
- Comparison of multiple models

Thank you

Ústav pro českou literaturu AV ČR
Institute of Czech Literature of the CAS

T E W