Topic Modelling 200 Years of Russian Drama

Irina Pavlova[1] · Frank Fischer[2,3]

[1]University of Oxford · [2]Higher School of Economics, Moscow · [3]DARIAH-EU

EADH 2018, Galway · 8th of December 2018

DraCor: Drama Corpora Platform

- https://dracor.org/ (public alpha!)
- an infrastructure for the research on European drama
- TEI corpora as basis (two in-house corpora: Russian and German drama corpora, 1730–1930)
- **DraCor API** to provide data for research questions
- facilitate access to specific text slices, e.g., for bag-of-words approaches like stylometry or topic modelling

What is topic modelling (TM)

In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract 'topics' that occur in a collection of documents. (Wikipedia 2018)

- the method is based on the co-occurrences of words in those documents
- the model presents topics as sets of frequently co-occurring words for extracted 'topics'
- the interpretation and naming of these 'topics' falls to the researcher

So many scholars in humanities departments are turning to [topic modelling] in their research that it is sometimes described as part of the digital humanities in itself. (Schmidt 2012)

Topic models are the mother of all collocation tools. (Jockers 2013, p. 123)

Goals

- build a suitable topic model for Russian Drama Corpus
  - establish appropriate parameters
  - introduce a method of less subjective evaluation of the model
- explore the data
  - apply the model to the data and interpret results for
    - topics’ temporal distribution
    - thematic differences between genres
    - topic tendencies in different authors

Related work

Matthew Jockers: *Macroanalysis*, *“Theme”*

- one of the first large attempts to topic model fiction (2013)
- > 3,000 English novels, 500 topics extracted
- *literature evolves partially – or even completely – independent of individual creativity* → topics reflect the general development of literature

- results divided by gender, nationality and time

Christof Schöch: *Topic Modelling Genre* (2017)

- French Drama of the Classical Age and the Enlightenment
- TM is a good approach for *discovering thematic patterns and trends in large collections of text*
- TM on fiction is more challenging: meanings and themes are often implicit in literary texts
- clusterisation by dramatic subgenres

Workflow

LDA

- two most popular algorithms: Non-negative matrix factorization (NMF) and Latent Dirichlet allocation (LDA)
- LDA is *a generative probabilistic model for collections of discrete data such as text corpora* (Blei et al. 2013)
- Python 3, scikit-learn

Size of documents

- plays?
- acts?
- one character’s speech-texts?

→ chunks

Stop-words

- standard stop-words set for Russian

  \+ characters’ proper names

  \- ambiguous terms, i.e. *матушка* or *батюшка*

Parts-of-speech

- POS restriction (only meaningful POS)
- Only nouns?
- Only verbs?

Other parameters

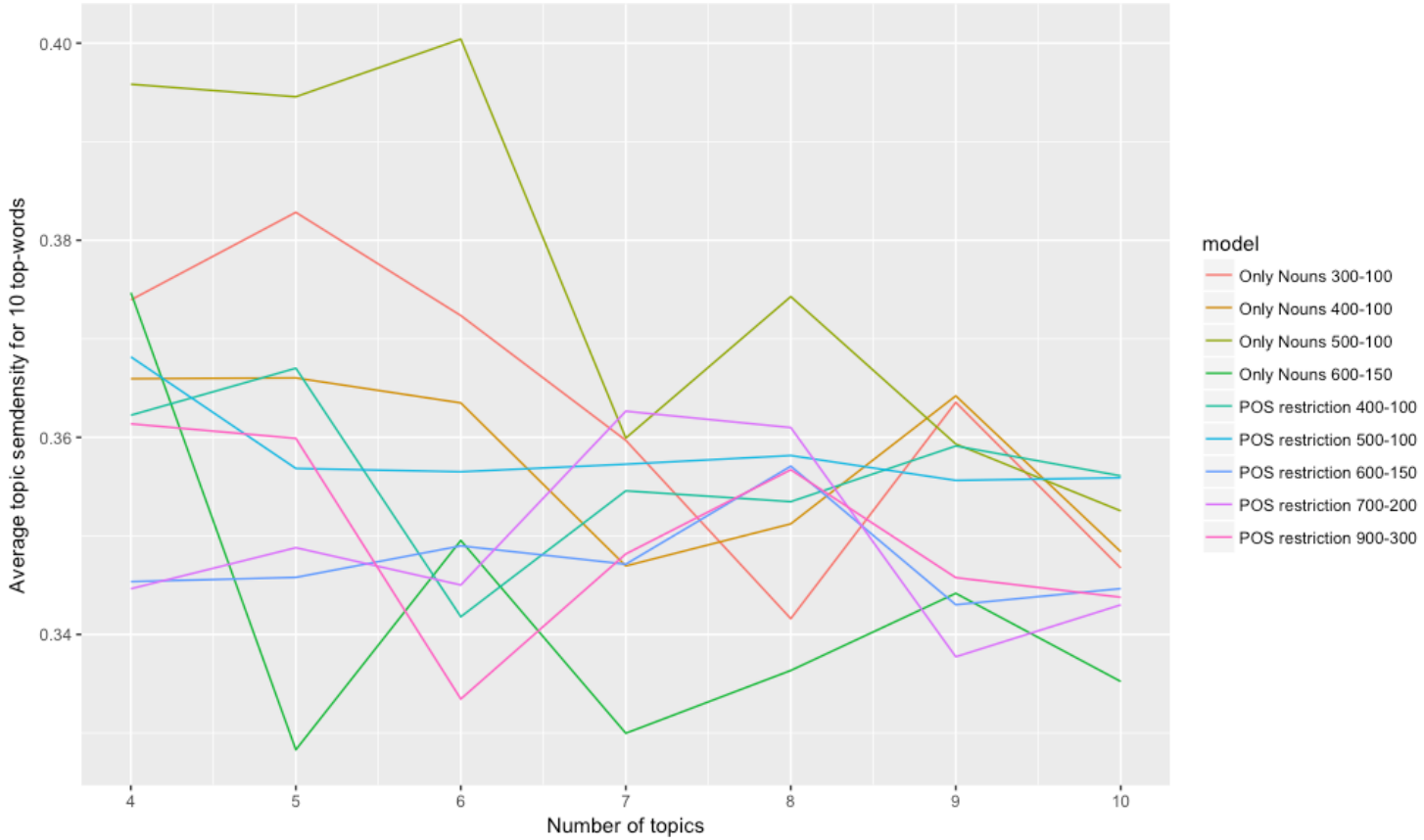
- num of iterations = 100
- *min_df* = 0.2
- *max_df* = 0.7

Choosing the best model

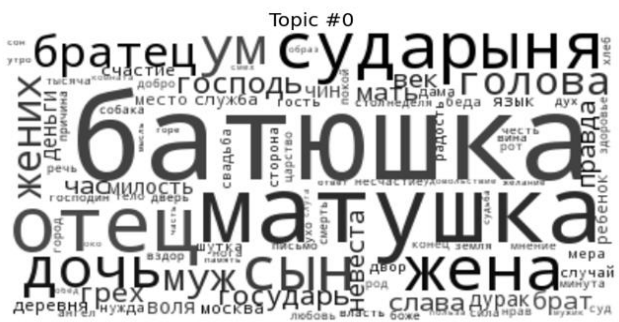
Semdensity

- distributional semantics
  - vectors based on Russian National Corpus (RusVectōrēs)
- measuring semantic density (semdensity) for different models
  - semantic closeness (cosine closeness of vectors) for a topic’s 10 top-words
  - average semantic closeness of all available pairs for 10 top-words for the topic
  - average semantic closeness of all topics in a model

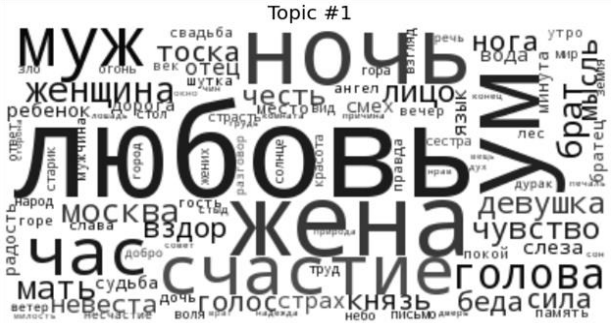
Semdensity


0.40
0.38
0.36
0.34
Average topic semdensity for 10 top-words
4
5
6
7
8
9
10
Number of topics
model
Only Nouns 300-100
Only Nouns 400-100
Only Nouns 500-100
Only Nouns 600-150
POS restriction 400-100
POS restriction 500-100
POS restriction 600-150
POS restriction 700-200
POS restriction 900-300

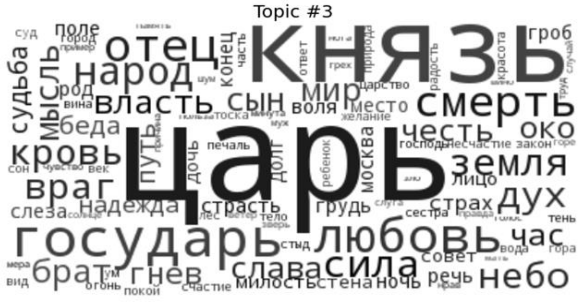
The topics



Family 1

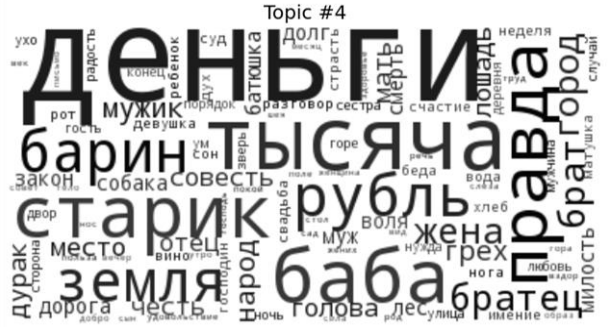


Love



Court

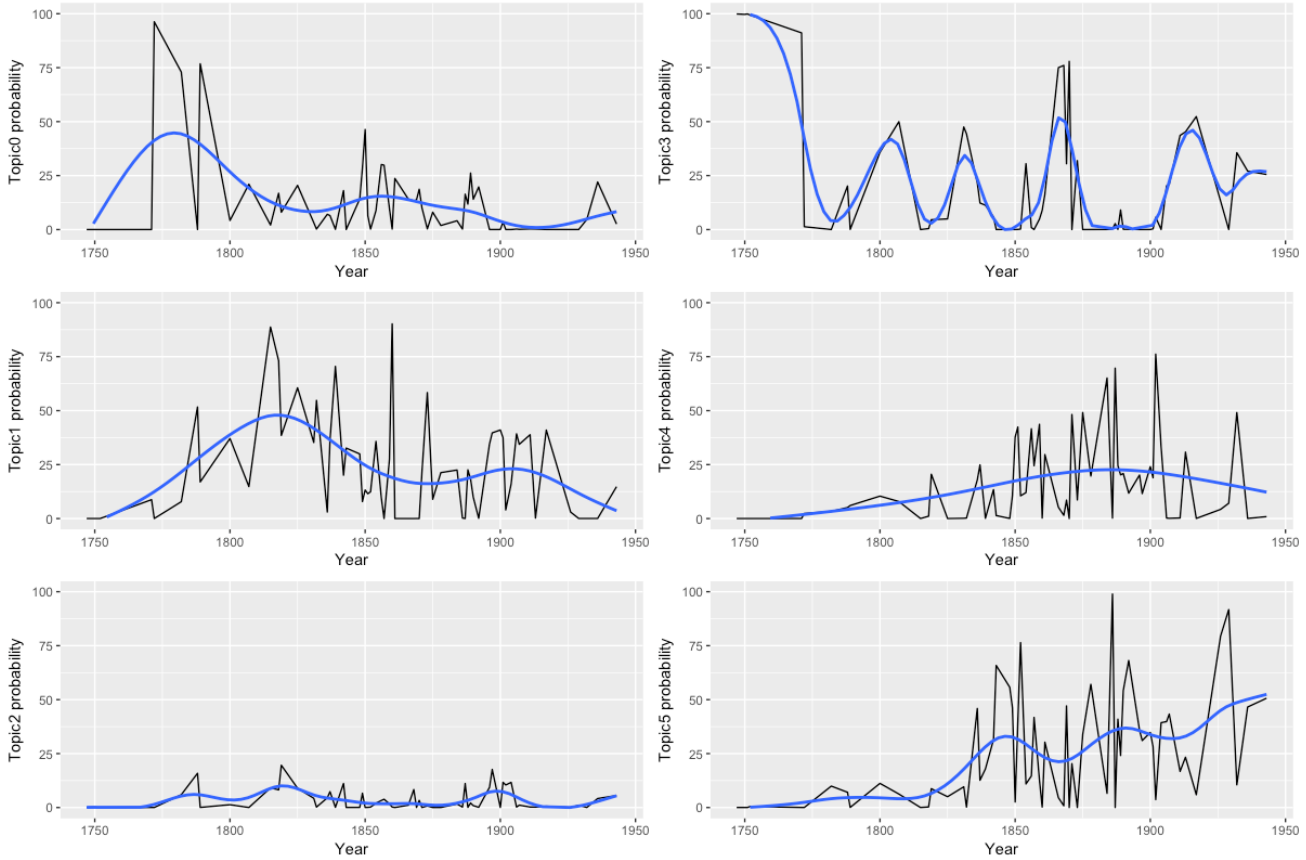
Family 2



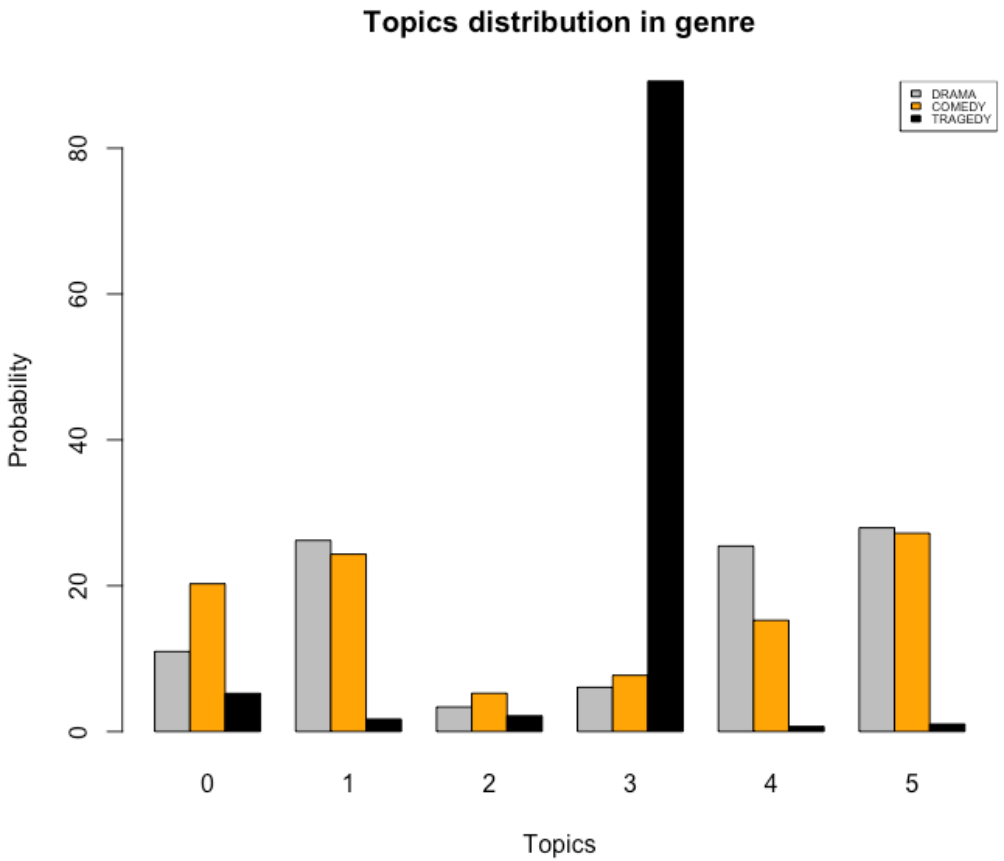
People

Results

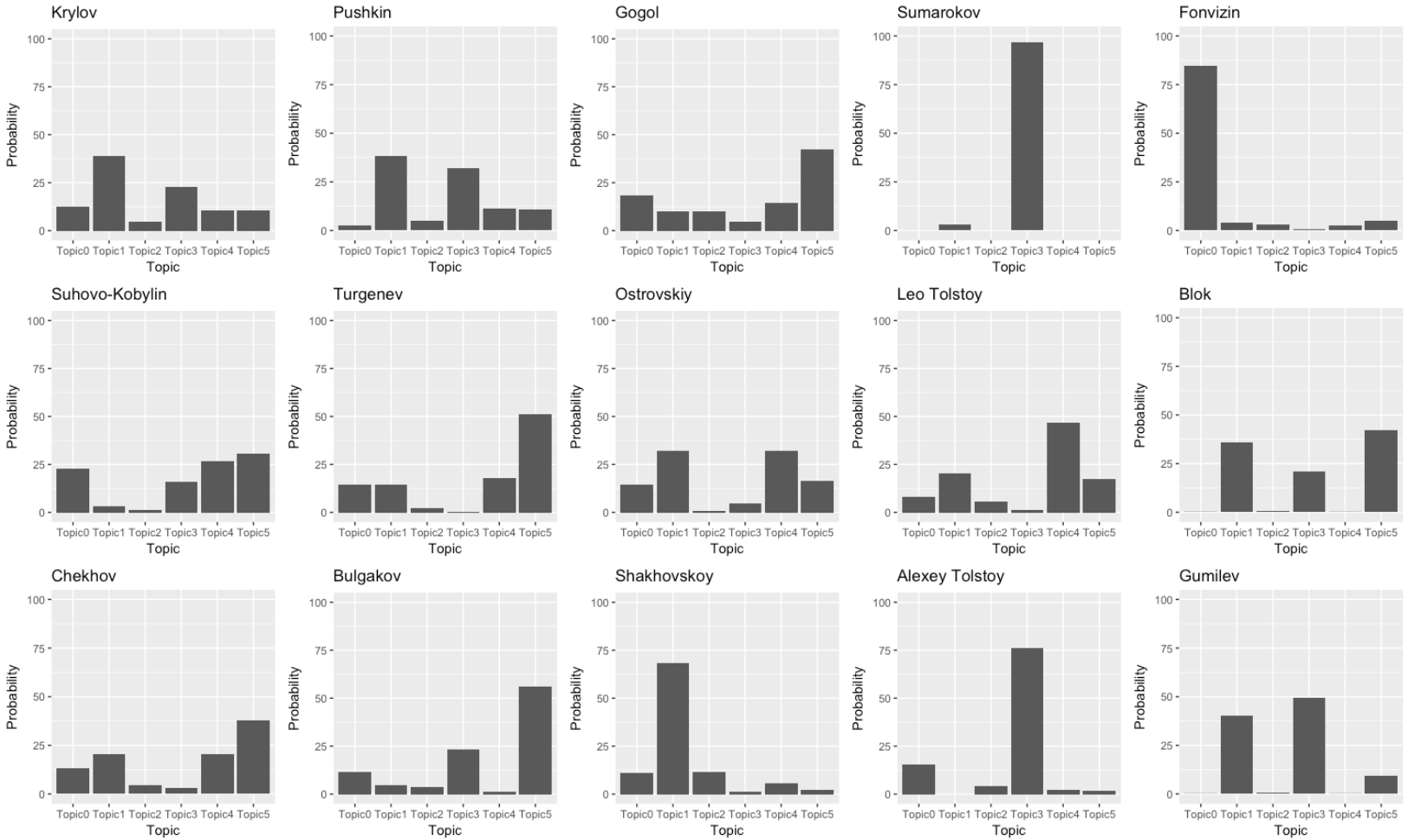
Temporal distribution



Genres


Topics distribution in genre
DRAMA
COMEDY
TRAGEDY
Probability
0
20
40
60
80
0
1
2
3
4
5
Topics

Authors


Krylov
Pushkin
Gogol
Sumarokov
Fonvizin
Suhovo-Kobylin
Turgenev
Ostrovskiy
Leo Tolstoy
Blok
Chekhov
Bulgakov
Shakhovskoy
Alexey Tolstoy
Gumilev
Probability
Topic
Topic0 Topic1 Topic2 Topic3 Topic4 Topic5
0 25 50 75 100

Conclusions and further work

- TM works on drama, including Russian drama
- mid-sized corpus can be enough for building a model
- Next stop: comparative study of topics in drama of different languages

Спасибо! Thanks!