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/](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

- 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
The topics
Results
Temporal distribution

![Graphs showing temporal distribution for topics 0, 1, 2, and 5.](image-url)
Genres

Topics distribution in genre

Probability

Topics

0 1 2 3 4 5
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!