Contemporary Russian Literature

Topic Modelling Methods

Ustinia Kosheleva, Anna Kondratjeva, Daria Maximova, Yevgeniy Lapin

Digital Humanities minor, Colloquium I, Feb. 18, 2017

Corpus

59 books (mainly novels and collections of short stories) by 12 contemporary Russian authors that were written and published in the years between 1984 and 2016.

List of authors: V. Pelevin, V. Sorokin, T. Tolstaya, D. Rubina, L. Ulitskaya, Z. Prilepin, Y. Vodolazkin, D. Bykov, M. Petrosyan, M. Veller, B. Akunin, L. Petrushevskaya.

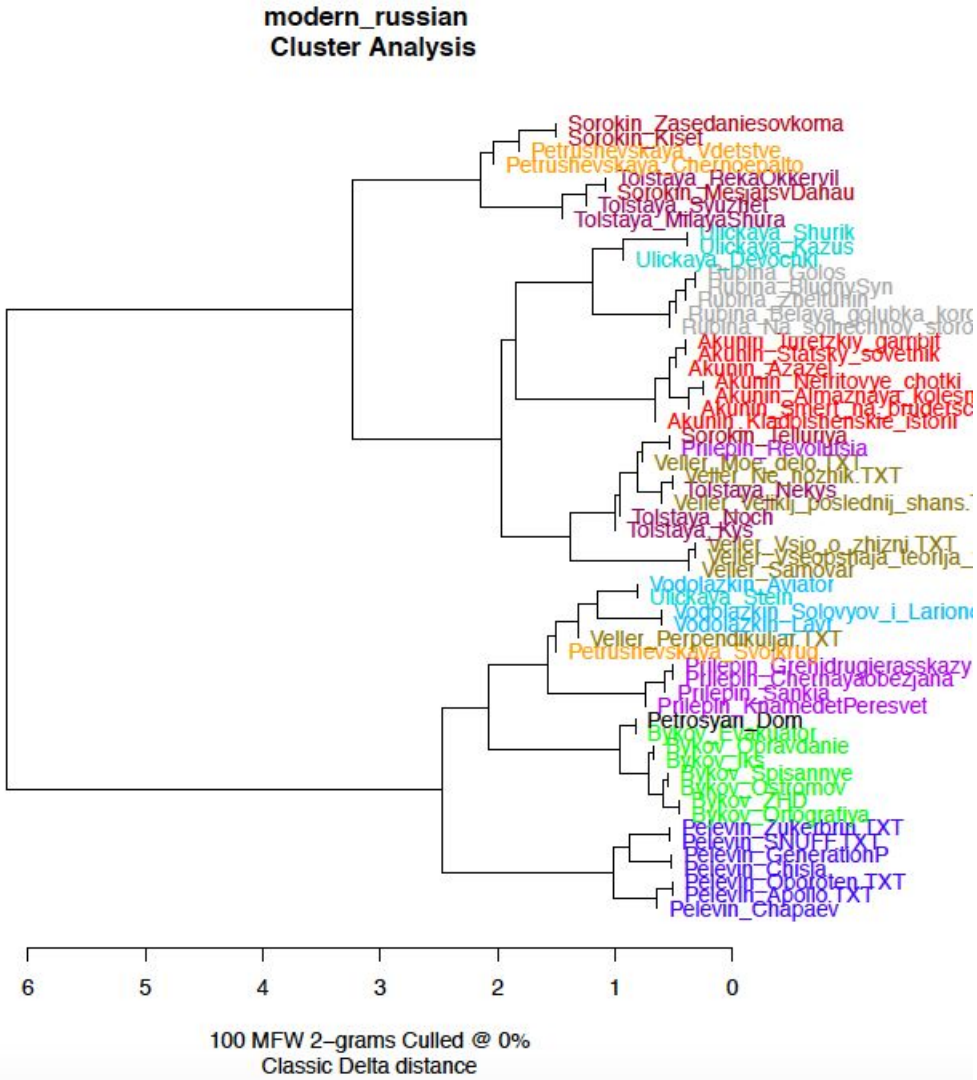
Full list may be found here.

Data preparation

- Annotations and notes were removed from the texts.
- Each book represented a ‘chunk’ of information, which was later processed with stylo() and mallet.
- In case of mallet texts were also tokenized and processed with Mystem to receive lemmas

(that’s what we were actually doing)

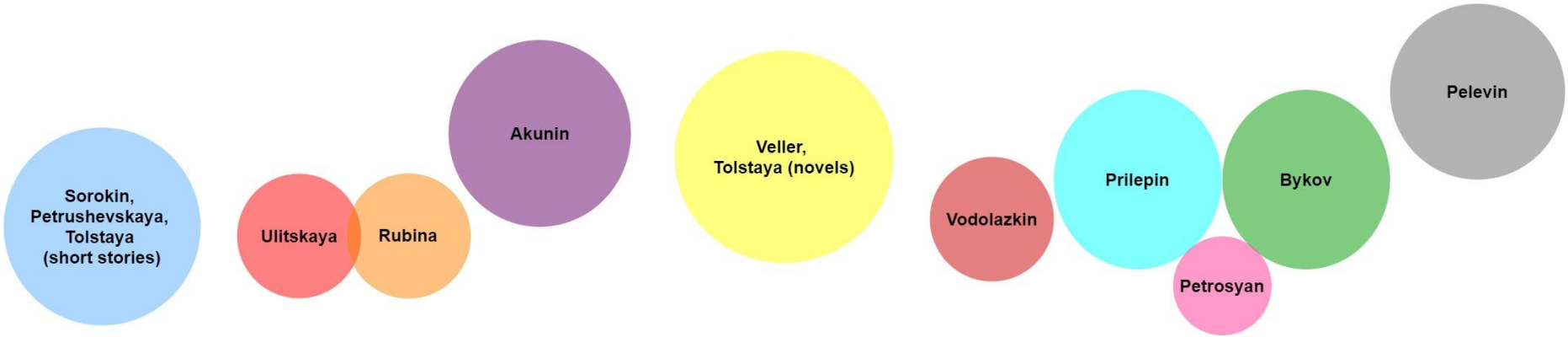
Stylometry


modern_russian
Cluster Analysis
Sorokin_Zasedaniesovkoma
Sorokin_Kiset
Petrushevskaya_Vdetstve
Petrushevskaya_Chernoepalto
Tolstaya_RekaOkkeryil
Sorokin_MesiatsvDanau
Tolstaya_Syuzhet
Tolstaya_MilayaShura
Ulickaya_Shurik
Ulickaya_Kazus
Ulickaya_Devochki
Rubina_Golos
Rubina_BludnySyn
Rubina_Zheltukhin
Rubina_Belaya_golubka_koro
Rubina_Na_solnechnoy_storo
Akunin_Turetzkiy_gambit
Akunin_Statsky_sovetnik
Akunin_Azazel
Akunin_Nefritovye_chotki
Akunin_Almaznaya_kolesn
Akunin_Smert_na_bruderso
Akunin_Kladbishenskie_istorii
Sorokin_Telluriya
Prilepin_Revolutsia
Veller_Moe_delo_TXT
Veller_Nie_hoznik.TXT
Tolstaya_Nekys
Veller_Vetkij_poslednij_shans.
Tolstaya_Noch
Tolstaya_Kys
Veller_Vsio_o_zhizni.TXT
Veller_Vseobshaja_teorija_
Veller_Samovar
Vodolazkin_Aviator
Ulickaya_Stein
Vodolazkin_Solovyov_i_Larion
Vodolazkin_Lavr
Veller_Perpendikuliar.TXT
Petrushevskaya_Svoikrugi
Prilepin_Grehidrugierasskazy
Prilepin_Chernayabezdana
Prilepin_Sankia
Prilepin_KhamedetPeresvet
Petrosyan_Dom
Bykov_Evakuator
Bykov_Opravdanie
Bykov_Iks
Bykov_Spisannye
Bykov_Ostromov
Bykov_ZHD
Bykov_Orfografiya
Pelevin_Zukerman.TXT
Pelevin_SNUFF.TXT
Pelevin_GenerationP
Pelevin_Chisla
Pelevin_Obokoten.TXT
Pelevin_Apollo.TXT
Pelevin_Chapaev
6
5
4
3
2
1
0
100 MFW 2–grams Culled @ 0%
Classic Delta distance

Stylo: problems

- Different times and genres of particular piece may be one of the reasons why some authors are not grouped together
- Corpus is not balanced out!
- But it seems that we are able to pick out some stylistic clusters depending on the results of few experiments.

Some kind of stylistic clusters



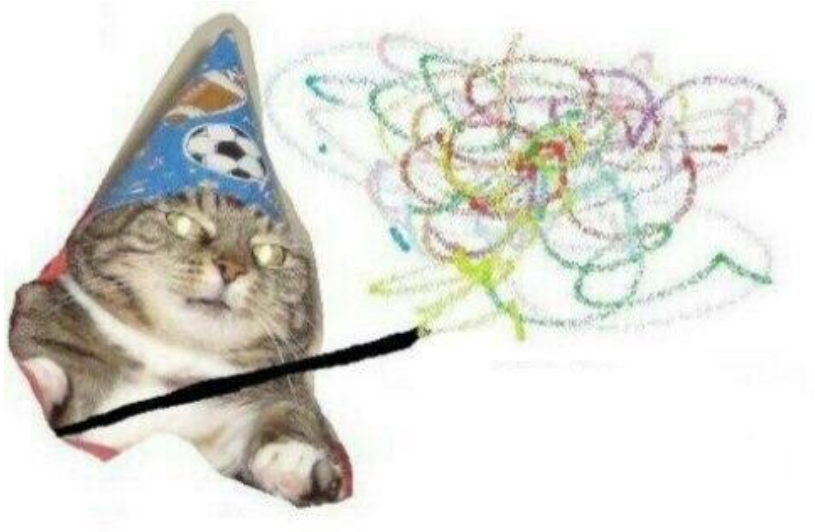
...so, it seems that if you like, for example, Prilepin's manner of writing, you should probably try Bykov too.

Mallet

Mallet: problems

Problems with getting the settings right:

- ➔ Do we have enough data?
- ➔ How do we get the number of iterations?
- ➔ How do we understand the topic of a group?
- ➔ How to edit the list of stop words?



Mallet: groups (at least we tried)

Below are the main (or the clearest) topics of our corpus according to Mallet's opinion (parameters: 15 topics, interval 20, 1000 iterations):

- русский литература писатель советский книга написать рассказ вообще слово герой хороший стр россия друг литературный автор фраза союз история поэт великий - **Literature** (Russian literature in particular)
- мир действие энергия сила ощущение вообще дело смысл счастье большой хороший общий сильный друг система природа уровень равный любовь возможность - **Creation**
- генерал товарищ офицер русский война армия поезд командир солдат дело военный штаб немец рука начальник ротмистр смерть - **War**

Mallet: groups (at least we tried)

Below are the main (or the clearest) topics of our corpus according to Mallet's opinion:

- господин дело советник сторона князь генерал русский мас титулярный плечо начальник отвечать дверь офицер минута агент полковник полиция - **State**
- минута последний ребенок отец большой дом мать женщина вечер друг улица утро комната час мама старый улыбаться маленький происходить квартира - **Family and Home**
- машина отвечать вставать город мужик москва русский хороший деревня губернатор дорога улица россия быстро пить водка жена последний парень стол - **...Russia?**

Overall conclusions and results:

1. ...there is still a lot of work to do.
2. But, at least, we can compare styles of different authors and get quite adequate results, even pick out some stylistical clusters.
3. And determine the clearest topics of corpus (it is debatable, but the results seem not so bad)

Yet, things to be done:

1. Balance out corpus
2. Even more narrow time frame
3. Figure out how the hell mallet does its job

Thank you for your attention!