Русский стилеметрический датасет

уроки и возможности

Борис Орехов

Назначение

Набор русскоязычных текстов, подготовленный для

- стилеметрических экспериментов,
- машинного обучения,
- учебных занятий по компьютерному анализу текста

Структура

Файлы сгруппированы по тематическим папкам

Статус

Все тексты находятся в общественном достоянии (Public Domain) и доступны для исследований без ограничений.

Временной охват

Произведения с 1763 по 1932 год (охватывает три века: XVIII, XIX и XX).

Тщательный подбор

Все тексты снабжены метаданными и сопоставимы по объему внутри коллекции

Состав датасета

01. Тексты

Более **320** документов

02. Слова

Совокупный объем — **16+ миллионов** слов

03. Авторы и подкорпуса

Тексты 94 авторов, распределенные по 25 подкорпусам

Разделы датасета

Имена файлов соответствуют шаблону автор_название.txt. Это позволяет пакету stylo (R) автоматически группировать тексты по классам при визуализации

01

По авторам

Художественная проза XVIII–XX вв., публицистика, научные труды

02

Спецвыборки

Сравнение авторов по полу, возрасту, работе с псевдонимами (Чехов / Чехонте)

03

Жанры и стили

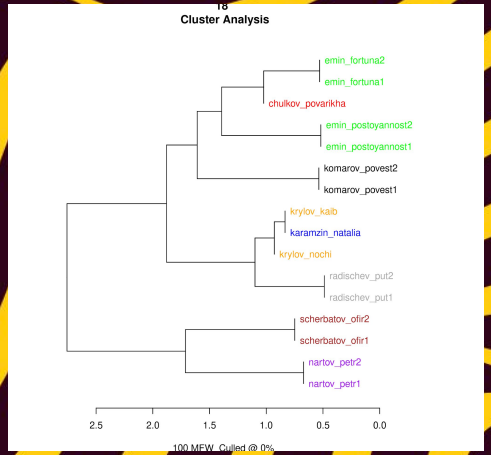
Романтизм vs Реализм, Западники vs Славянофилы, стихи vs проза

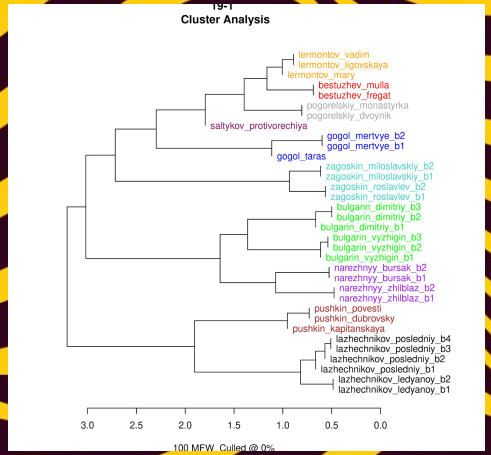
04

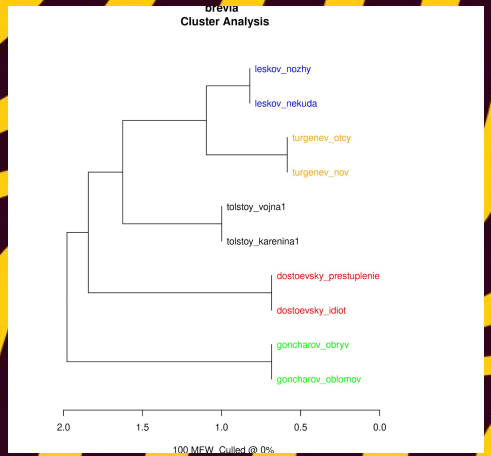
Специфические

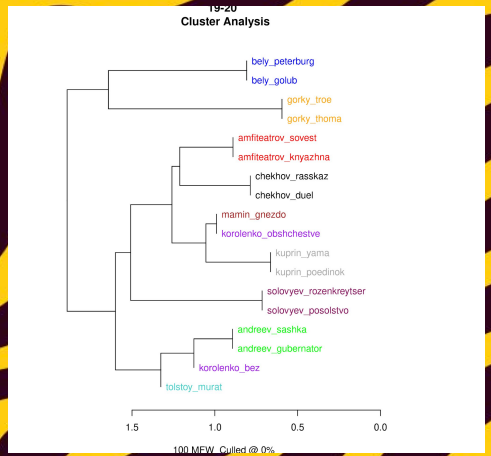
Речь персонажей и тексты, сгенерированные нейросетями

Разведочный анализ: Delta, 100 MFW









Авторы художественной литературы

Почти без сбоев

Не в свои кластеры попали только Карамзин и Короленко из 34 авторов

Volume 38, Issue 3
September 2023

Article Contents

Abstract

Hacking stylometry with multiple voices: Imaginary writers can override authorial signal in Delta

Get access >

Daniil Skorinkin ✉, Boris Orekhov

Digital Scholarship in the Humanities, Volume 38, Issue 3, Septem 2023, Pages 1247–1266, https://doi.org/10.1093/llc/fqad012
Published: 08 April 2023

Cite Permissions Share ▾

Abstract

VIEWS ALTMETRIC

430

More metrics information

Email alerts

New journal articles

Activity related to this article

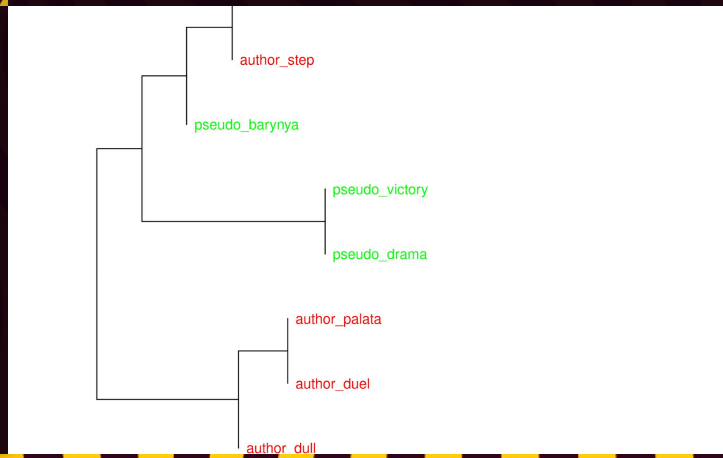
Автор и псевдоним

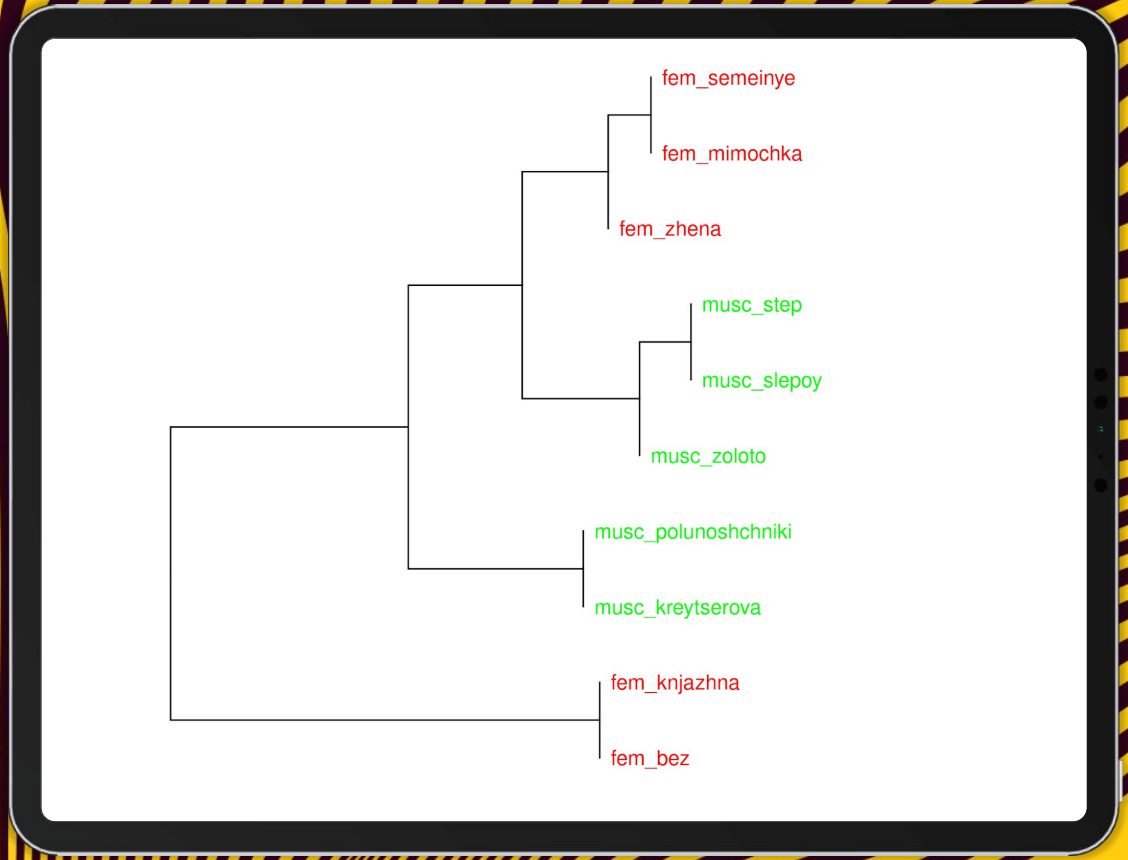
Чехов и Чехонте
Тексты под псевдонимом и под реальной фамилией

Граница: начало 80-х
Чехов переходит на практику публикации под своей фамилией

Псевдоним
Драма на охоте, Цветы запоздалые, Ненужная победа

Фамилия
Степь, Дуэль, Палата № 6

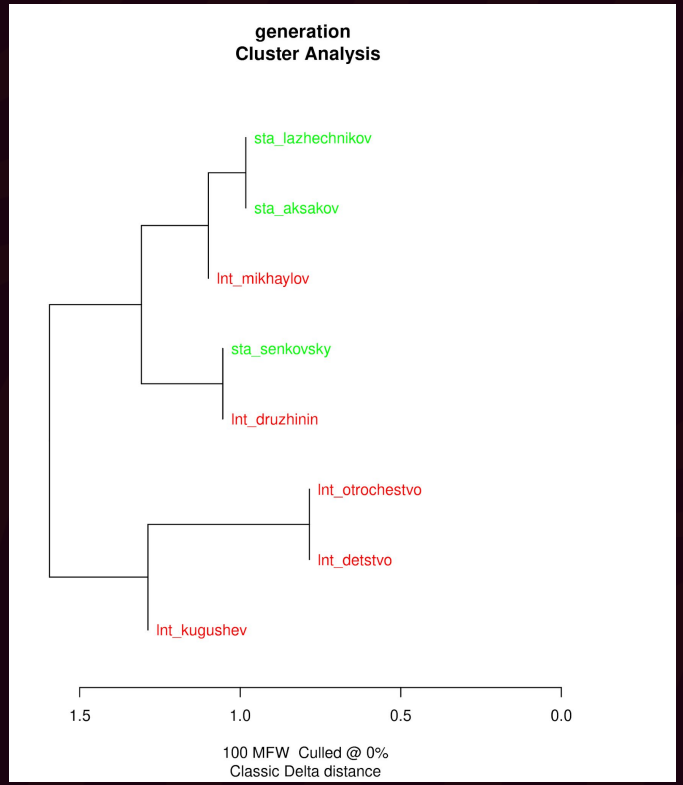




Пол автора

Тексты 1880–1890-х годов

Не идеально, но обнадеживающе



Поколения

Аксаков и Толстой

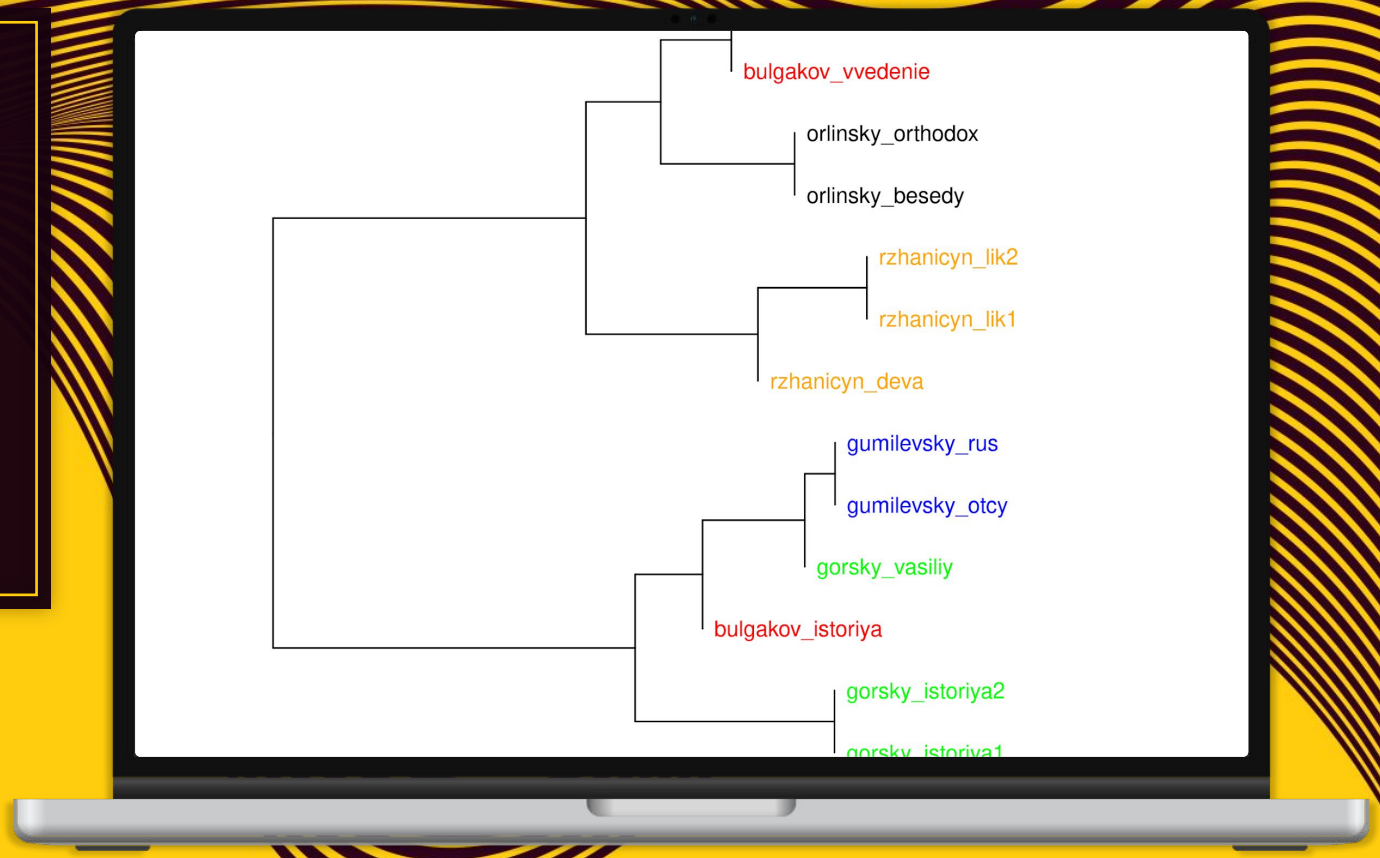
Лажечников, Сенковский (1790-е) vs. Дружинин и др. (1820-е)

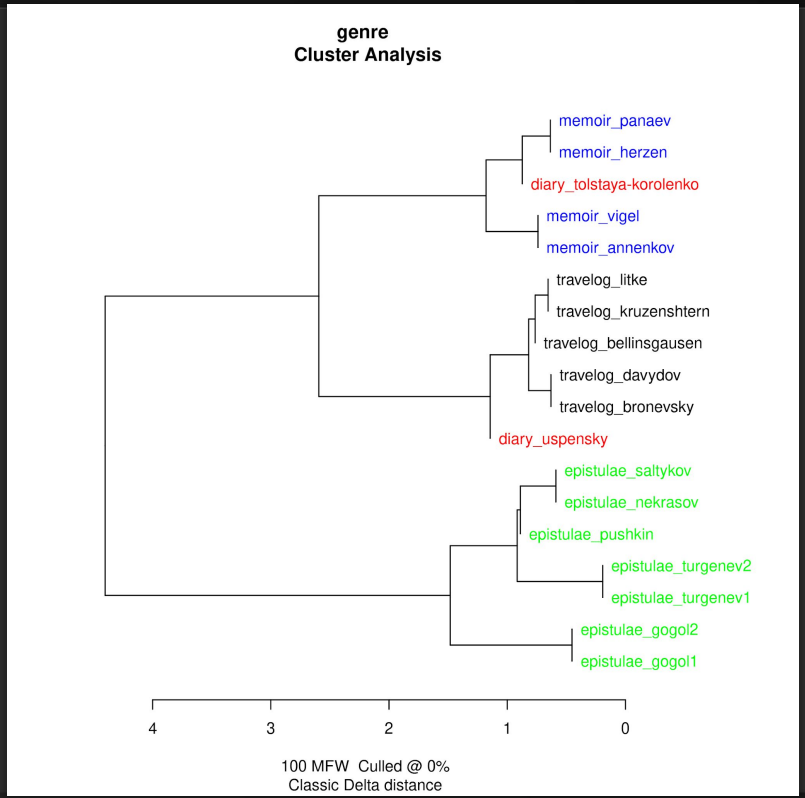
1856—1859

«Семейная хроника», «Детство»

Теология

Очень непоследовательные результаты
Возможно, из-за библейских цитат

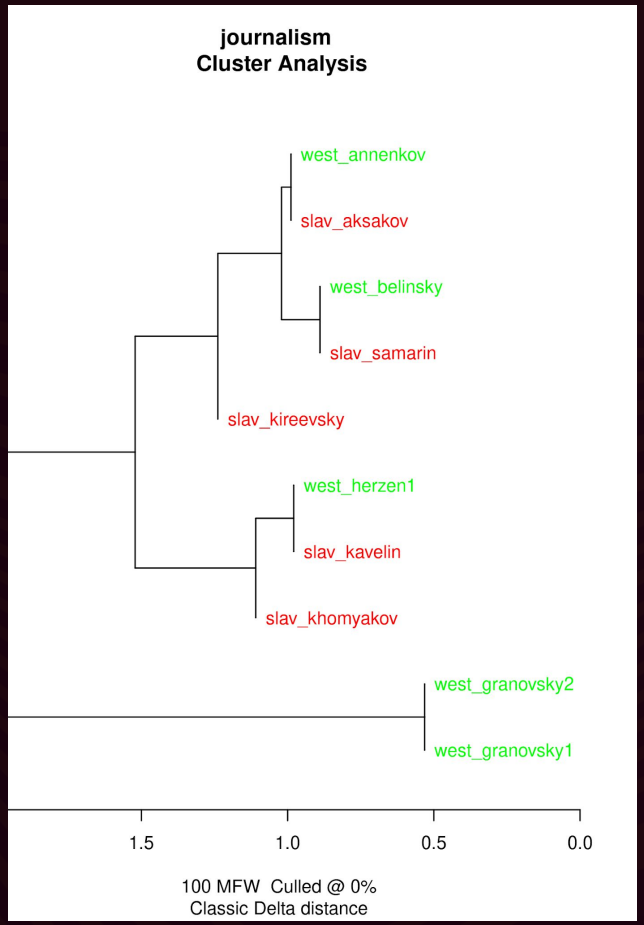




Жанровый сигнал

Size matters

Дневники ~10 тыс слов; остальные от 80 до 200 тыс

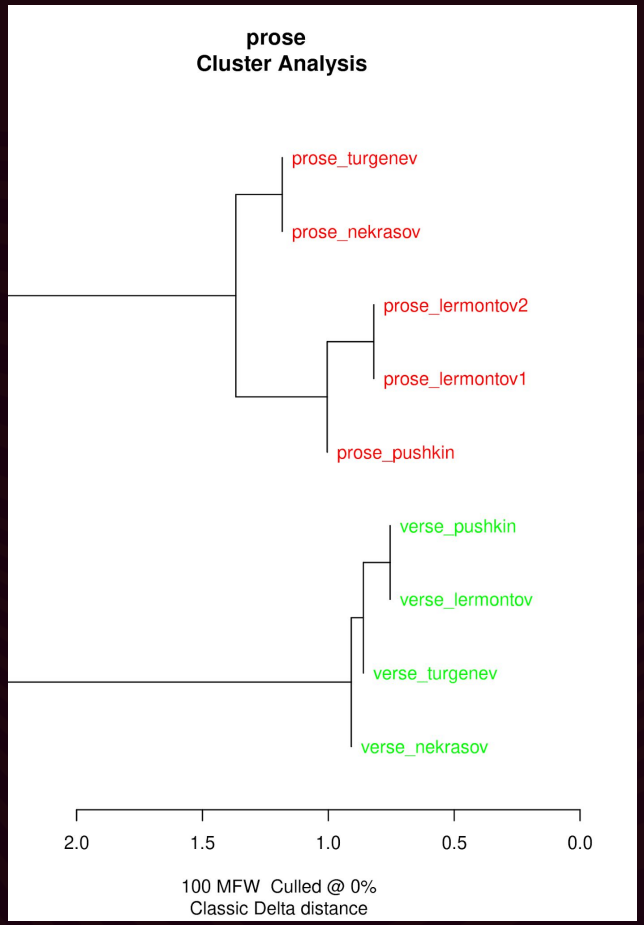


Западники славянофилы

Идейные течения

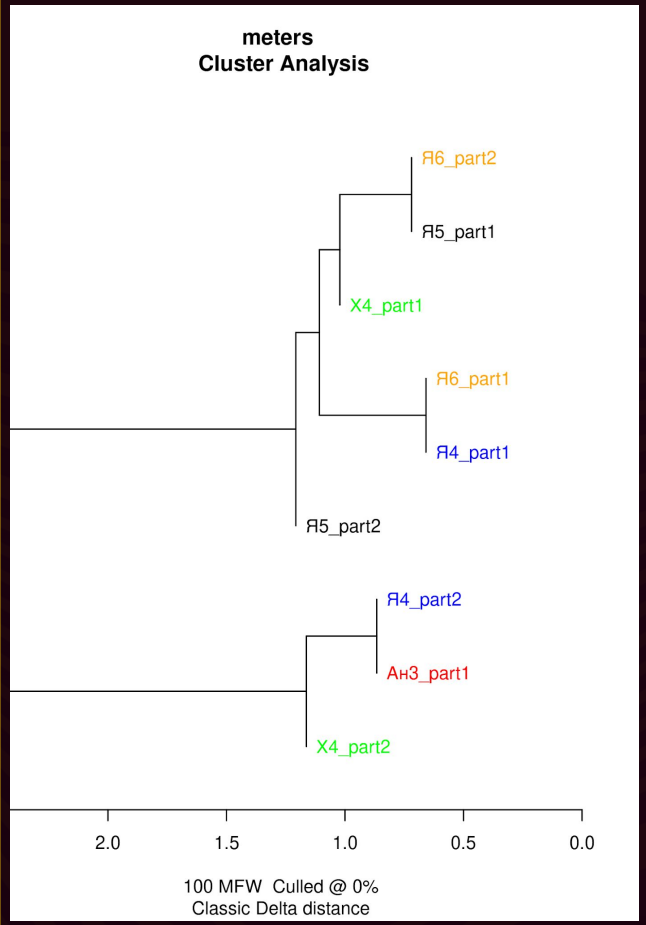
Кажется, нет

Умеет ли Delta отличать идеологию?



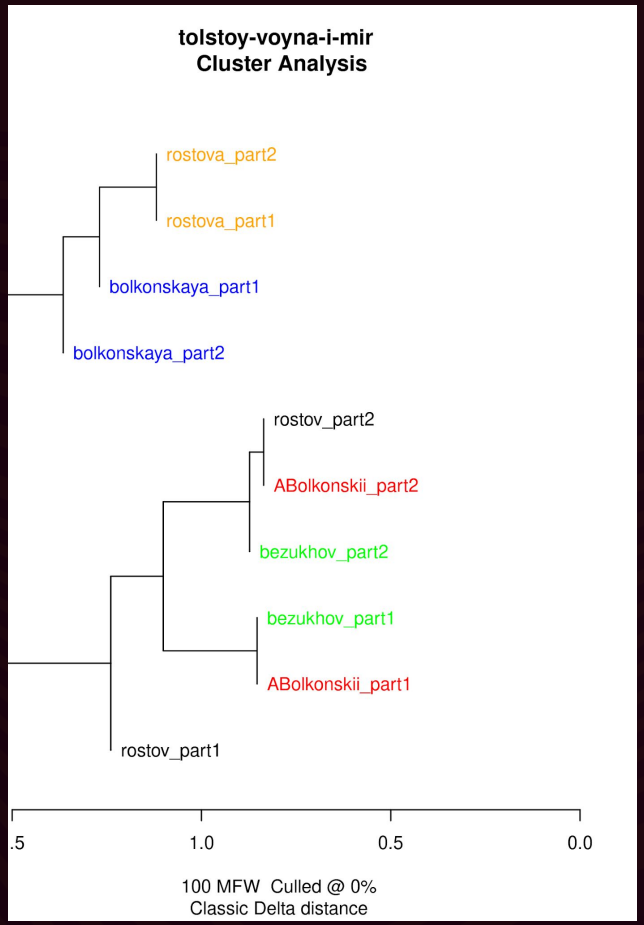
Стих и проза

Самый явный сигнал



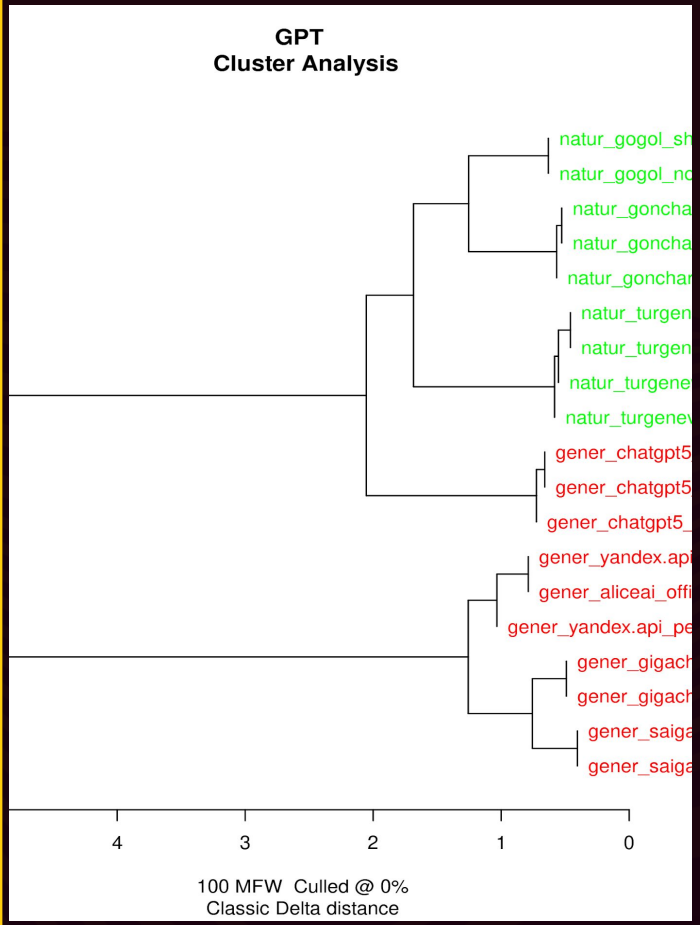
Поэтические размеры

Не различаются



Персонажи «Войны и мира»

К Д. А. Скоринкину есть вопросы



Естественные и сгенерированные

Перспективно

Спасибо!