Введение в цифровые гуманитарные исследования

Борис Орехов

Организационное

План курса

1. Лекция (31.10)
2. Практика (14.11)
3. Практика (5.12)
4. Конференция (12.12)
5. Конференция (19.12)

Полезная ссылка

Материалы курса: https://nevmenandr.github.io/portfolio/teaching/msu2026



Теоретическая рамка

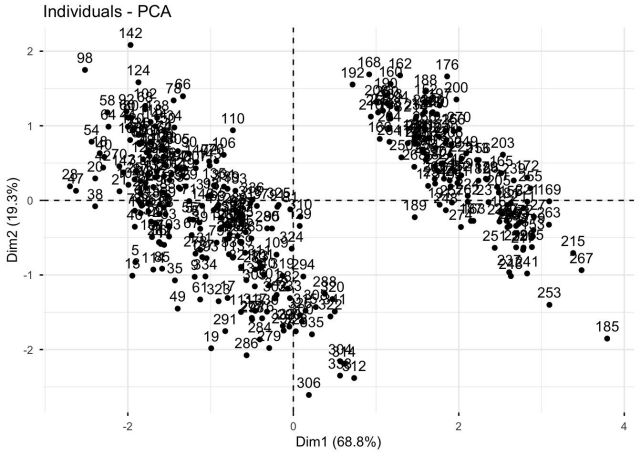
Зачем это нужно?

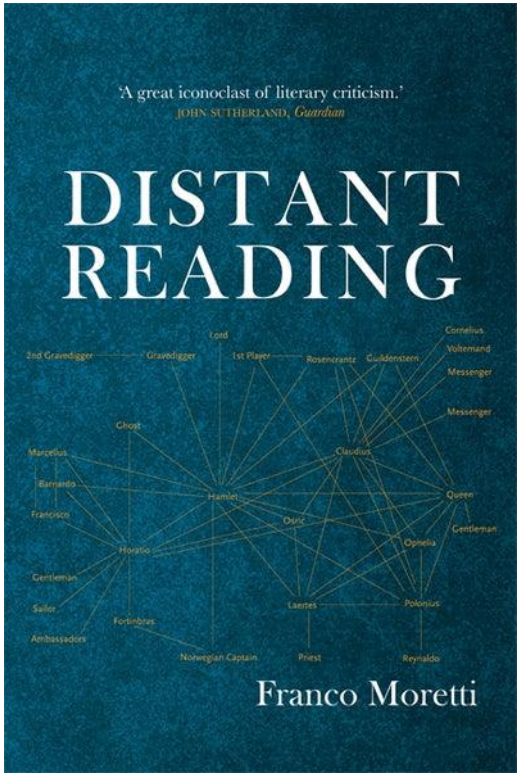
- Человечество живет в ситуации перепроизводста текстов.
- Формируется Великое Непрочтенное.
- Все это может прочитать компьютер.
- Distant reading.

Distant reading

https://vk.ru/wall-222835606_397 — дискуссия о переводе Моретти на русский 10 лет спустя.

Суть метода: между текстом и исследователем появляется дополнительная инстанция





Лучше всего работает

1. С атомарными фактами
   - отдельными словами,
   - предложениями,
   - именованными сущностями
2. С большими объемами

Примеры исследований

Текстовые коллекции

Культуромика

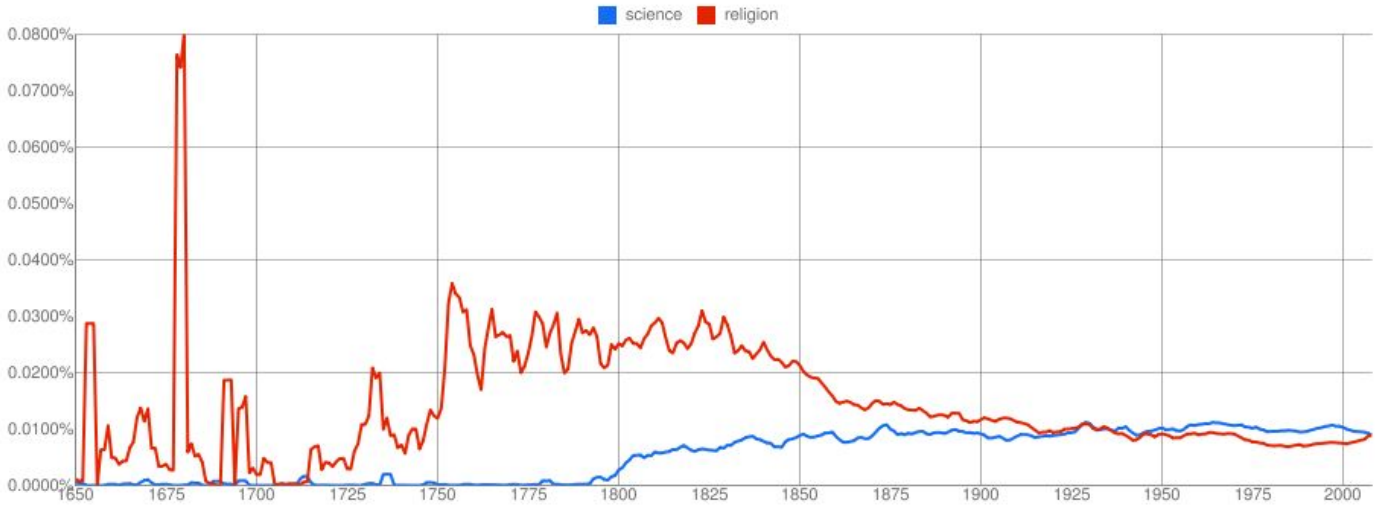
Методика основывается на гипотезе, что частотность слов в книгах отражает культурные тенденции.

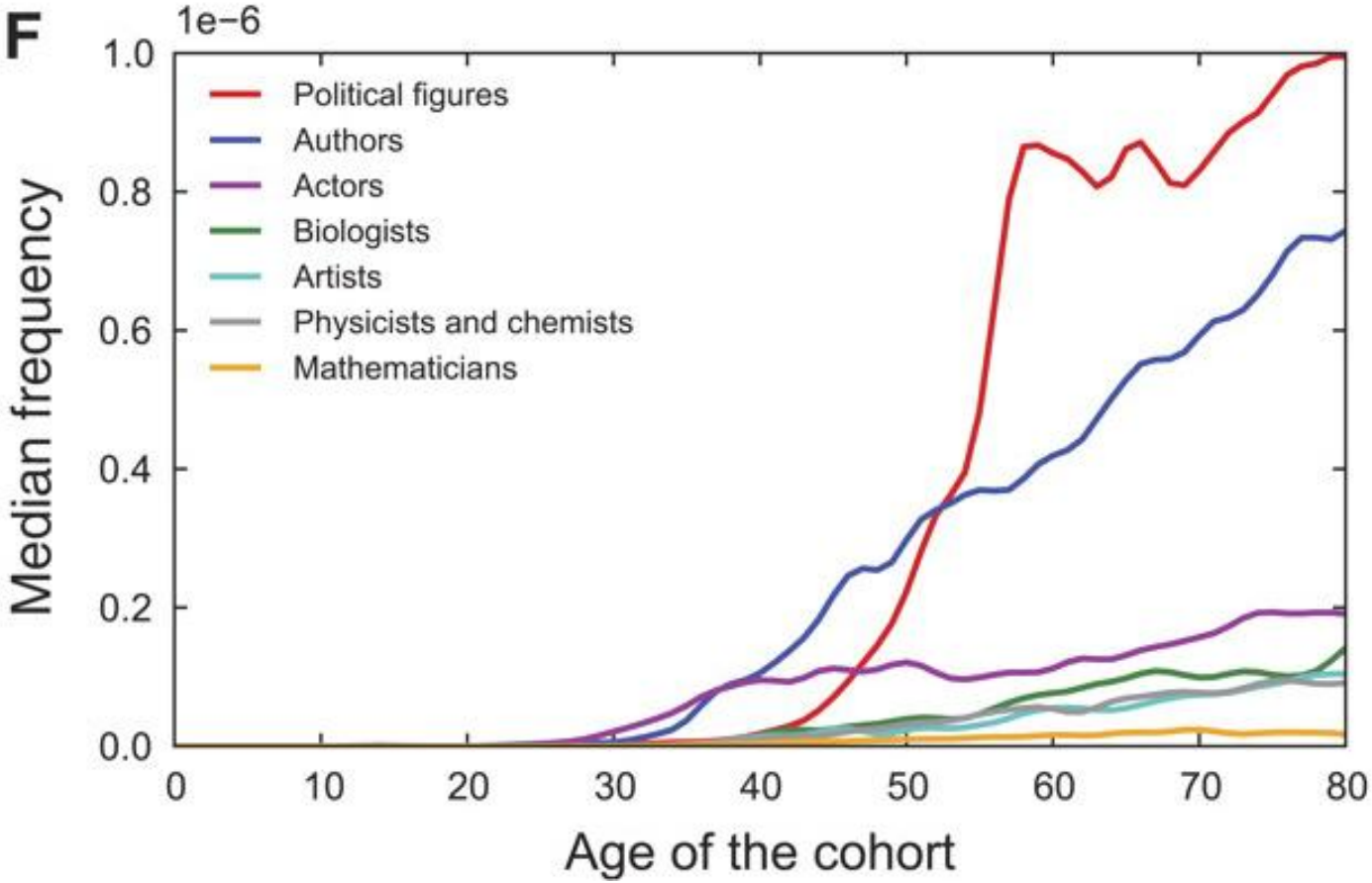
Инструмент: Google Ngram Viewer, миллионы книг, печатные издания

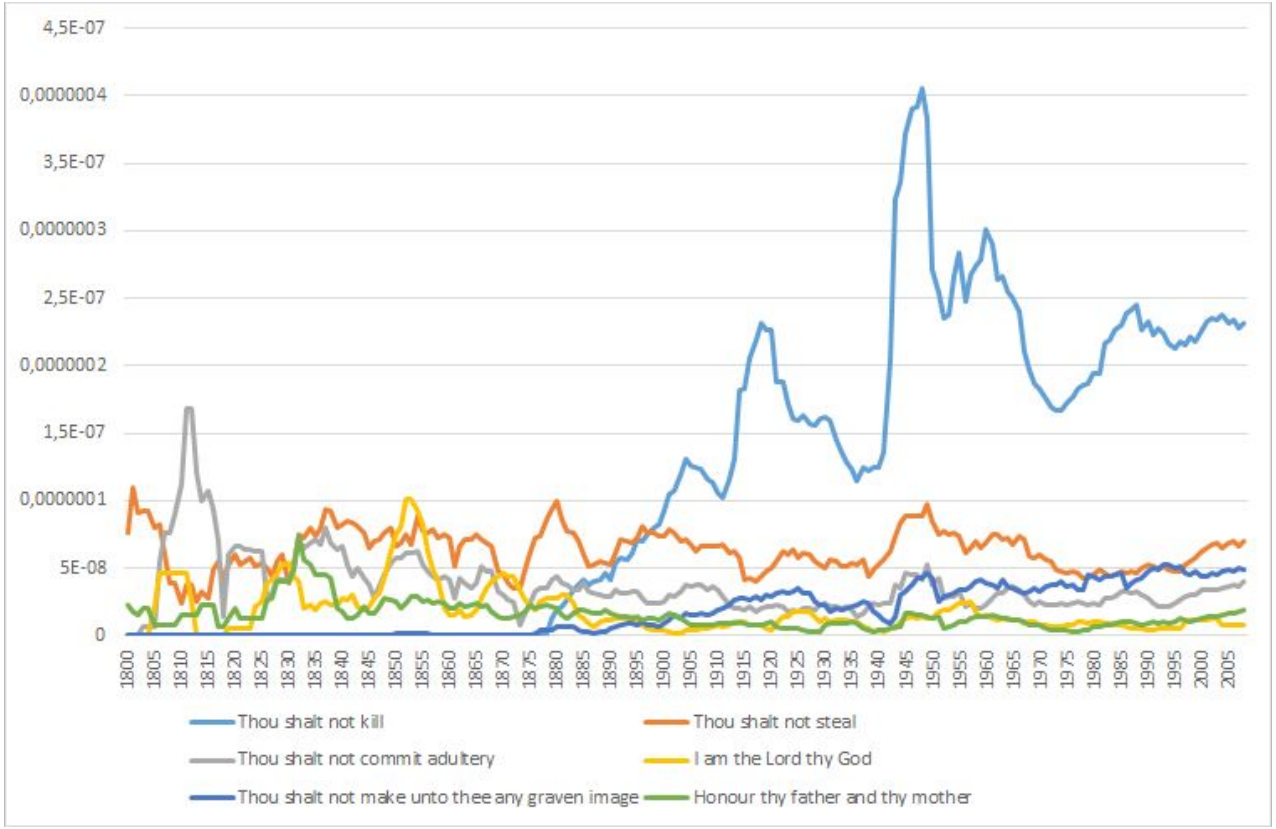
с 1470-х по 2000 г. Michel, J.-B., Liberman A., Aiden A. P., Veres A., Gray M. K., Pickett J. P., Hoiberg D., Clancy D., Norvig P., Orwan J., Nowak M., Pinker St.Quantitative Analysis of Culture Using Millions of Digitized Books // Science. 16 December 2010. 331 (6014): 176 – 82.

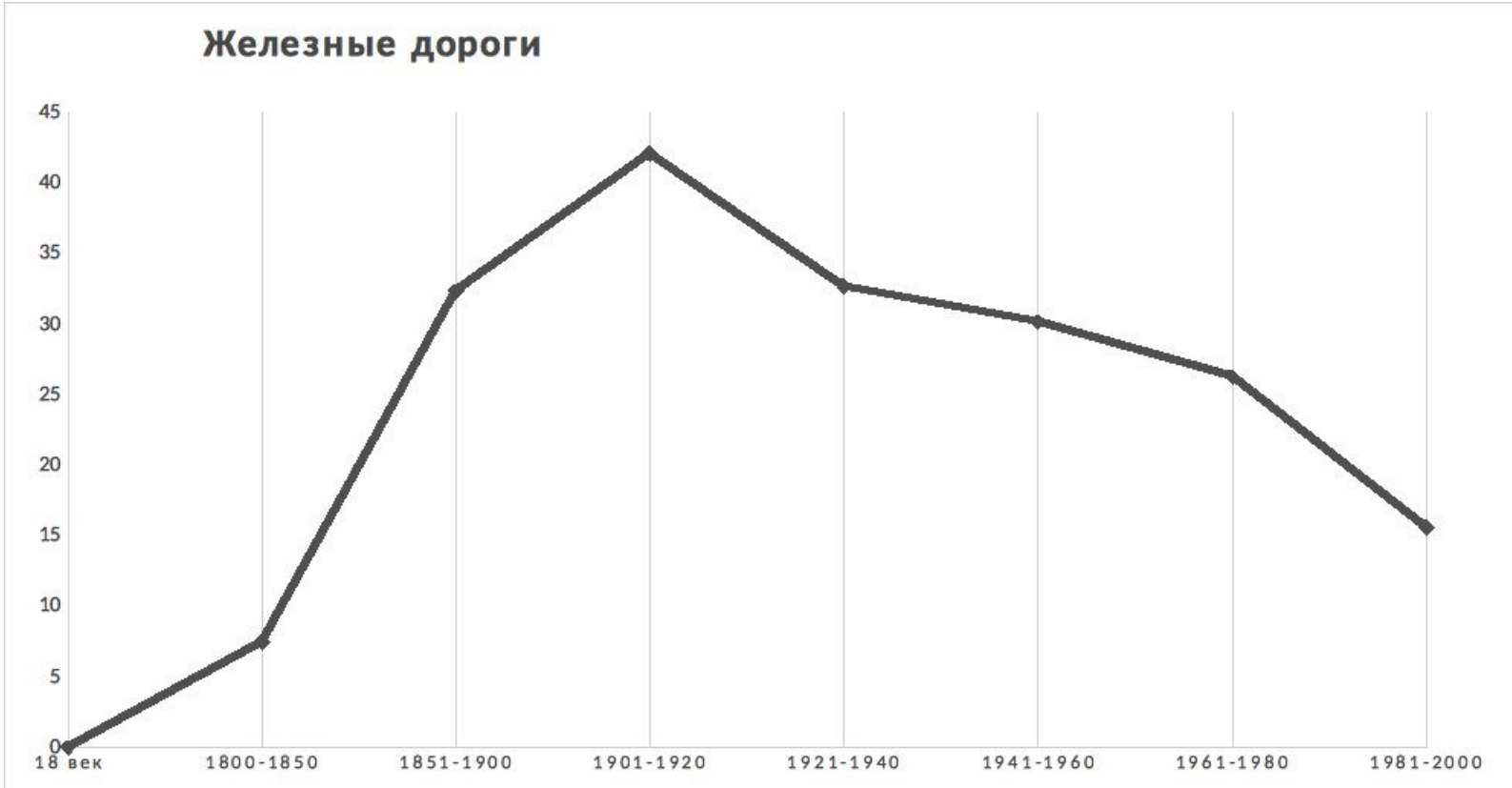
Google labs Books Ngram Viewer

Graph these **case-sensitive** comma-separated phrases: science,religion
between 1650 and 2008 from the corpus English with smoothing of 1 .
Search lots of books

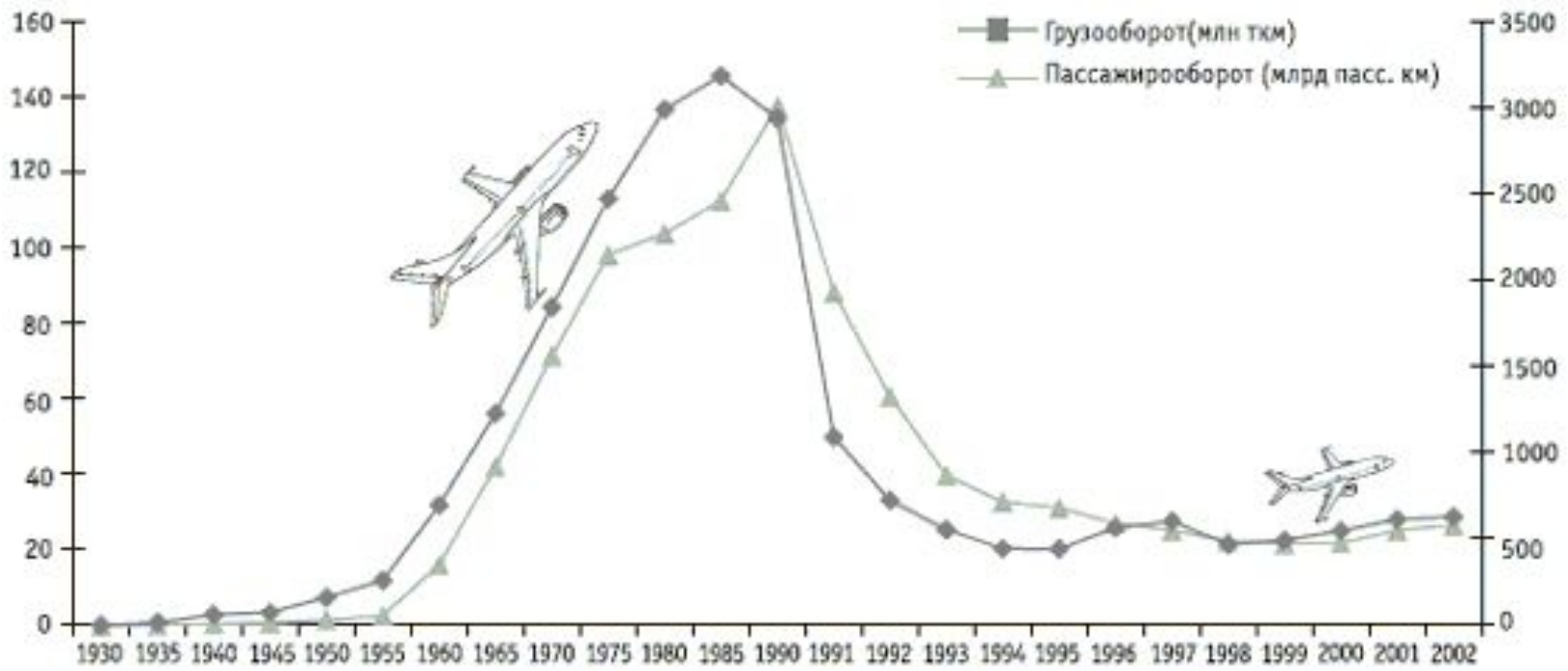



F
1e−6
1.0
0.8
0.6
0.4
0.2
0.0
Median frequency
Political figures
Authors
Actors
Biologists
Artists
Physicists and chemists
Mathematicians
0
10
20
30
40
50
60
70
80
Age of the cohort

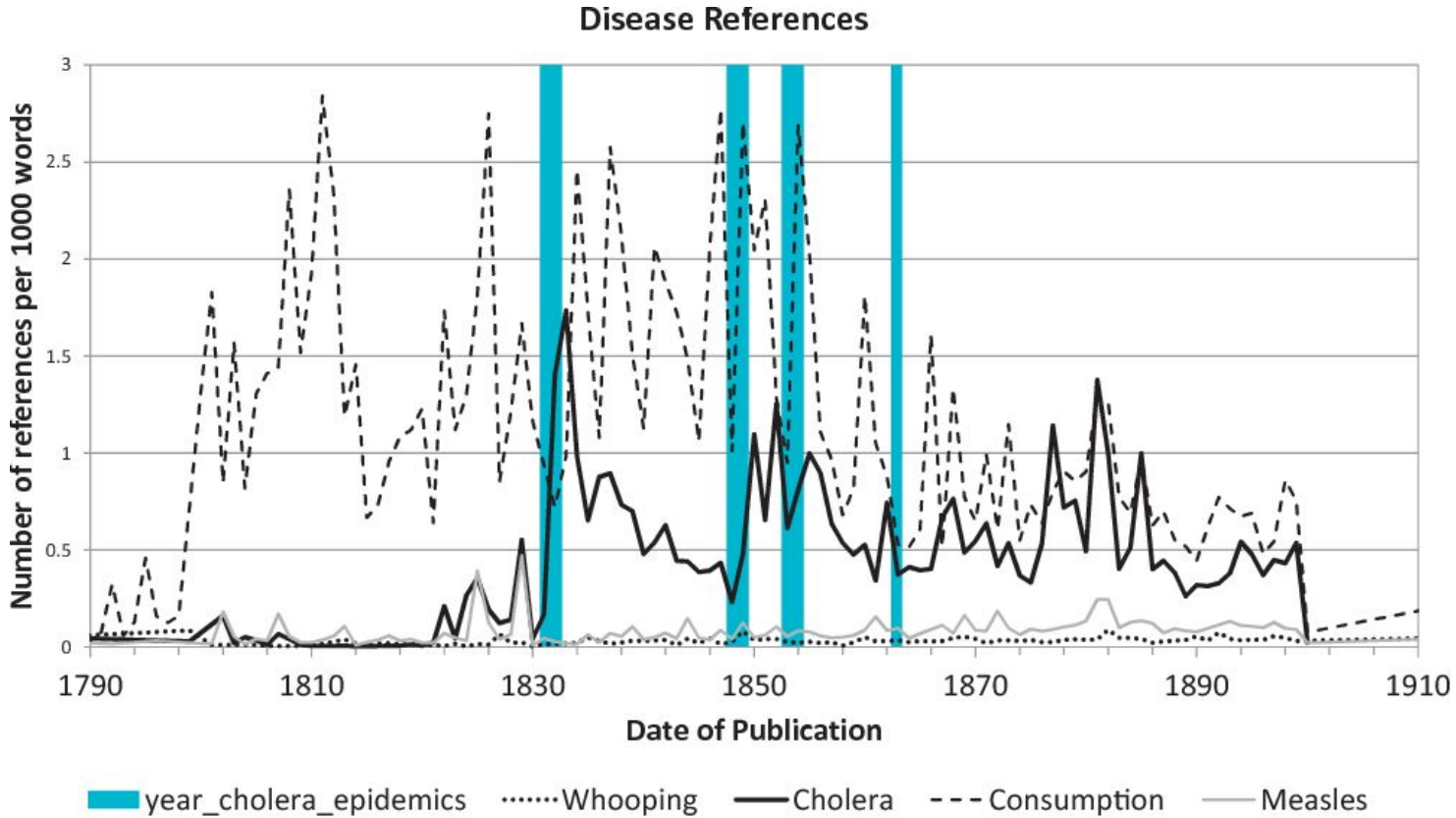

4,5E-07
0,0000004
3,5E-07
0,0000003
2,5E-07
0,0000002
1,5E-07
0,0000001
5E-08
0
1800
1805
1810
1815
1820
1825
1830
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1900
1905
1910
1915
1920
1925
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
Thou shalt not kill
Thou shalt not steal
Thou shalt not commit adultery
I am the Lord thy God
Thou shalt not make unto thee any graven image
Honour thy father and thy mother



Бонч-Осмоловская А. А. Культуромика в национальном корпусе русского языка, к постановке задачи: три века русских дорог



Бонч-Осмоловская А. А. Культуромика в национальном корпусе русского языка, к постановке задачи: три века русских дорог



Terras, Melissa & Baker, James & Hetherington, James & Beavan, David & Austwick, Martin & Welsh, Anne & O'Neill, Helen & Finley, Will & Duke-Williams, Oliver & Farquhar, Adam. (2017). Enabling complex analysis of large-scale digital collections: humanities research, high-performance computing, and transforming access to British Library digital collections. Digital Scholarship in the Humanities. 33. 456-466. 10.1093/llc/fqx020.

Корпусный анализ

Орехов Б. В. «Проблеск» Ф. И. Тютчева в ретроспективе Корпуса. Очерк корпусной поэтики // Корпусный анализ русского стиха: Сборник научных статей. Вып. 2 / Отв ред. В. А Плунгян, Л. Л. Шестакова. — М: Издательский центр «Азбуковник», 2014. — С. 305–319.

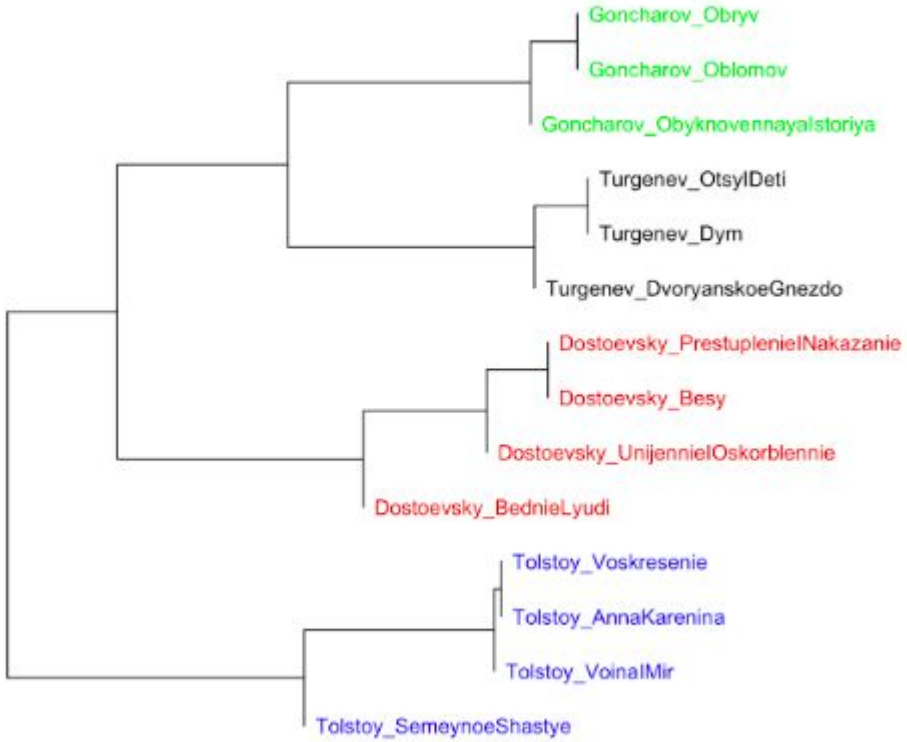
Проблеск

Слыхал ли *в сумраке* *глубоком*
Воздушной арфы легкий **звон**,
Когда полуночь, ***ненароком***,
Дремавших струн **встревожит** сон?..

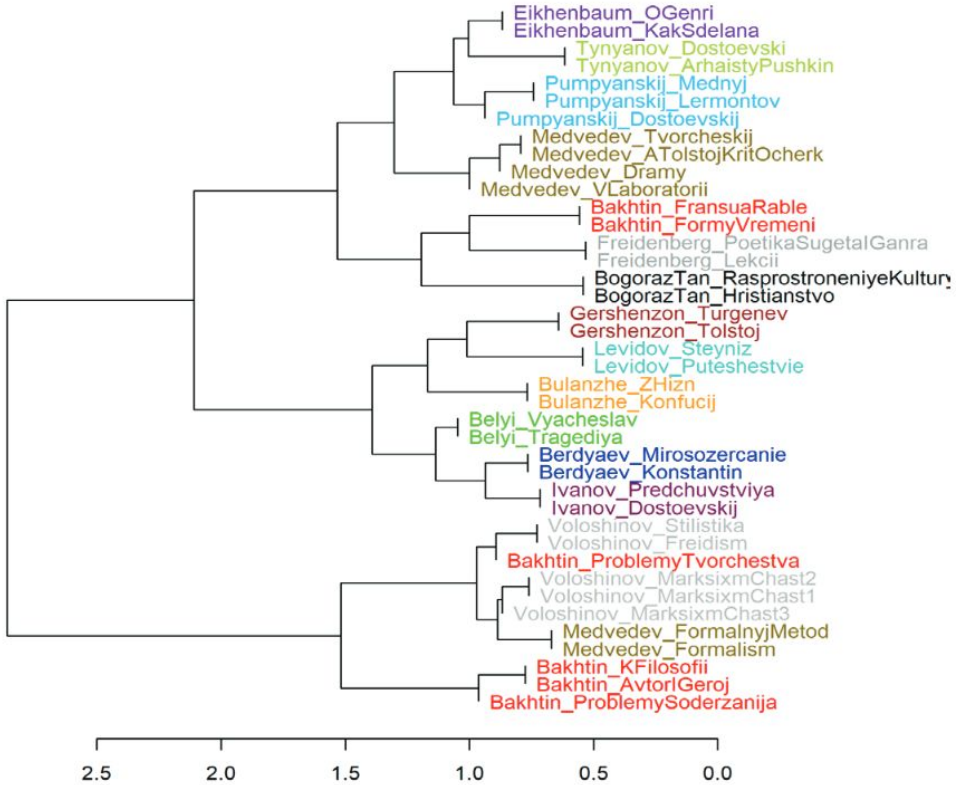
То потрясающие звуки,
То замирающие **вдруг**...
Как бы последний ропот муки,
В них отозвавшися, **потух**!

Дыханье каждое **Зефира**
Взрывает скорбь в ее струнах...
Ты скажешь: ангельская лира
Грустит, **в пыли**, по небесах!

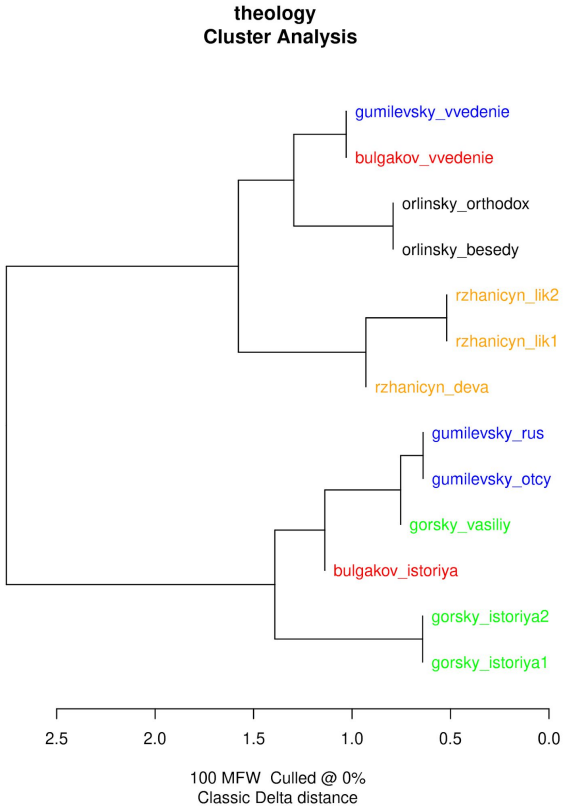
Стил(о|е)метрия



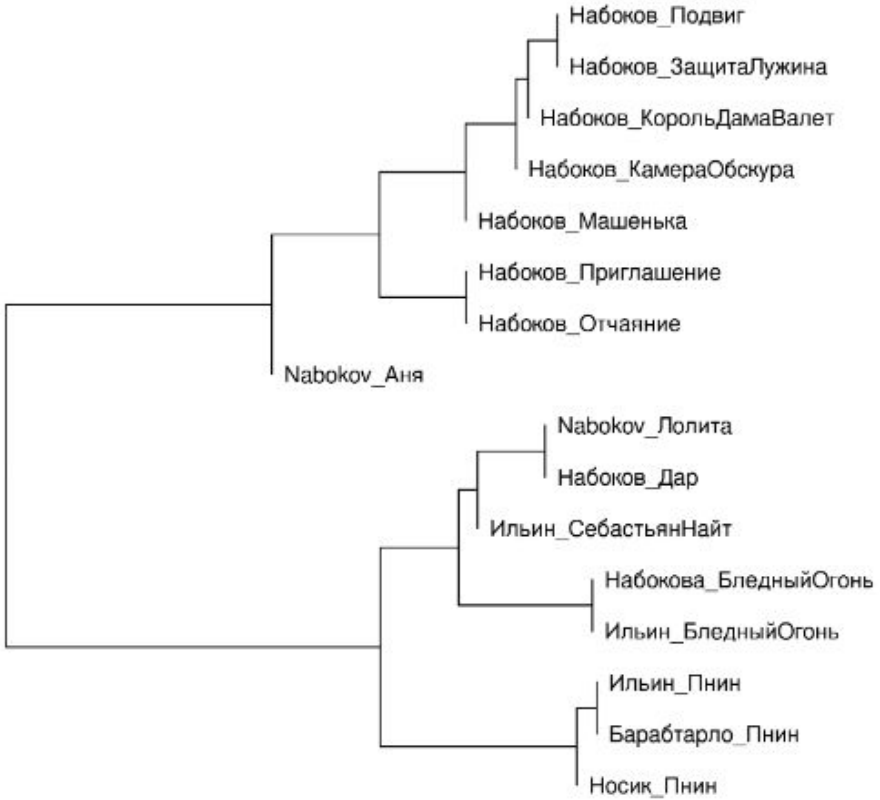
https://github.com/nevmenandr/RSD



Орехов Б. В., Василенко А. Г. Количественная атрибуция и «бахтинский круг»: проблема авторства в интеллектуальных сообществах // Новый филологический вестник. — 2026. — № 1. — С. 40–55. DOI: 10.54770/20729316-2026-1-40



https://github.com/nevmenandr/RSD/tree/main/author/nonfiction/science/theology

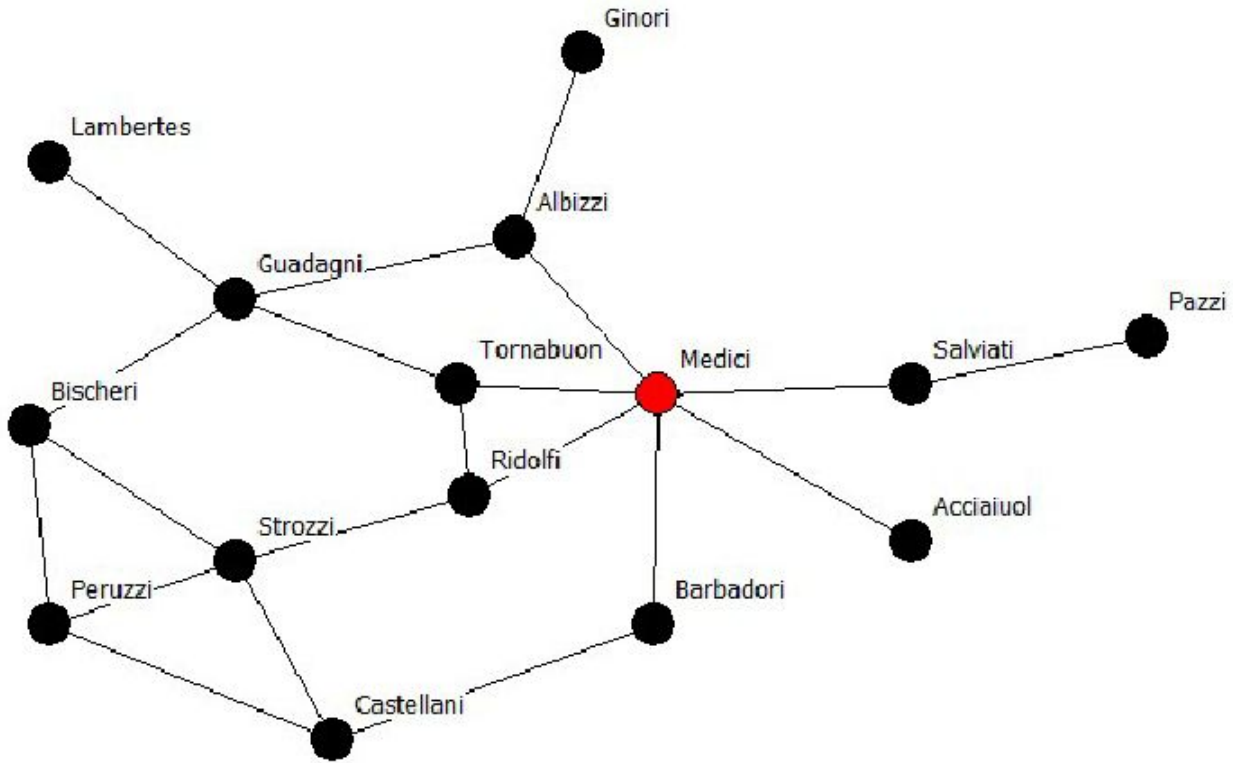


Орехов Б. В. Текст и перевод Владимира Набокова через призму стилеметрии // Новый филологический вестник. — 2021. — № 3. — С. 200–213.

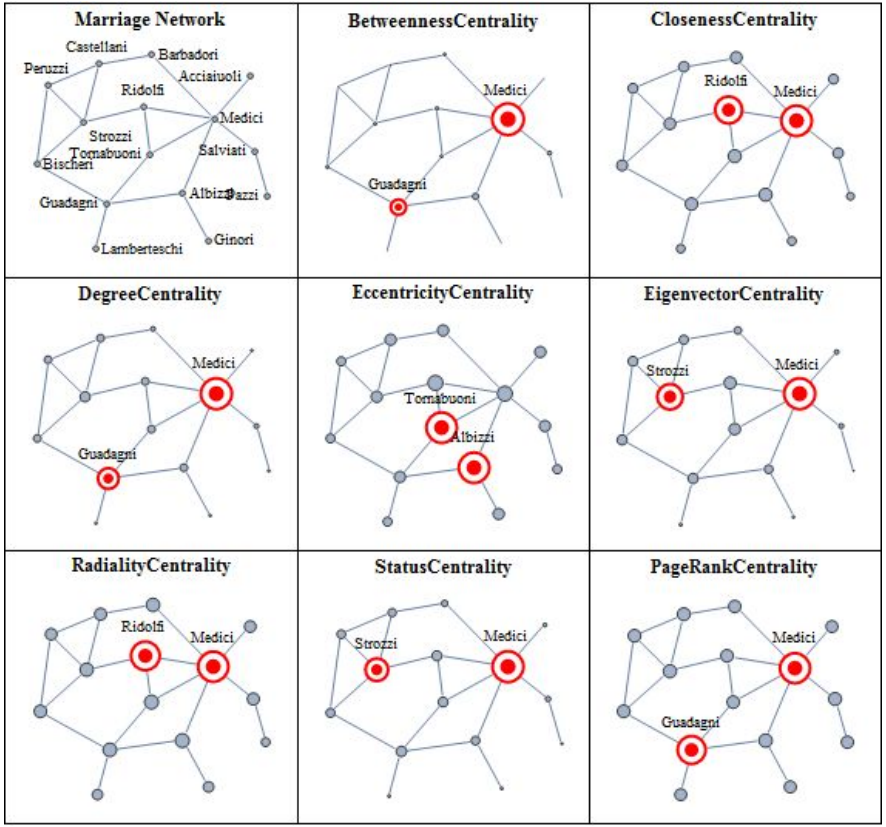
Сети

Сеть и сетевой анализ

- Сеть это модель, в которой самое важное это отношения между объектами.
- Сеть — это математический объект.
- Существуют алгоритмы, которые позволяют вычислить самые важные элементы сети.
- Существуют метрики оценки сети и отдельных ее элементов.
- Существуют алгоритмы, позволяющие выделить сообщества внутри такой сети.



Kent D. The rise of the Medici: Faction in Florence, 1426—1434. — Oxford: Oxford University Press, 1978.



Kent D. The rise of the Medici: Faction in Florence, 1426—1434. — Oxford: Oxford University Press, 1978.

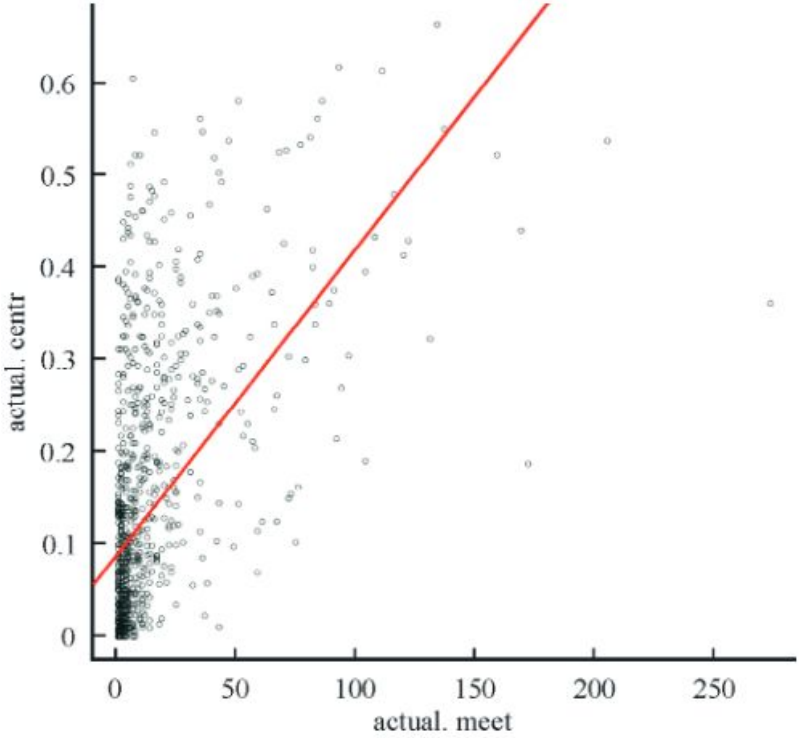
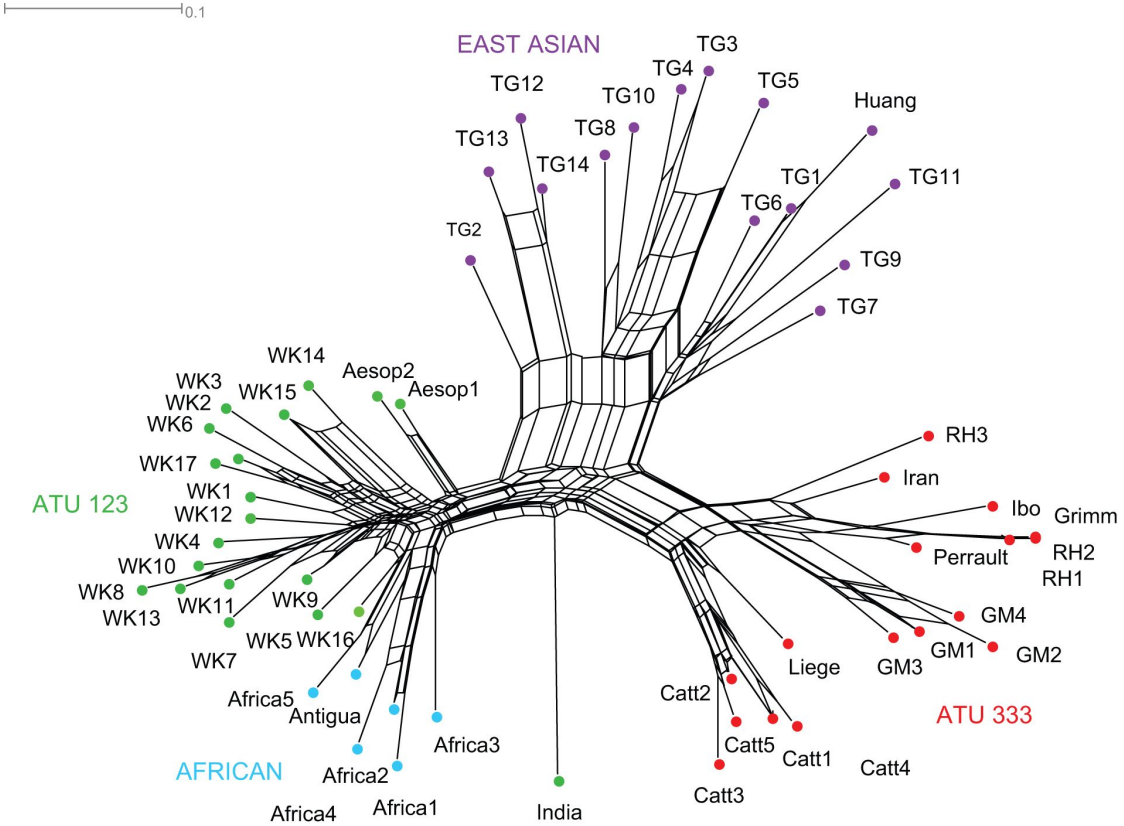


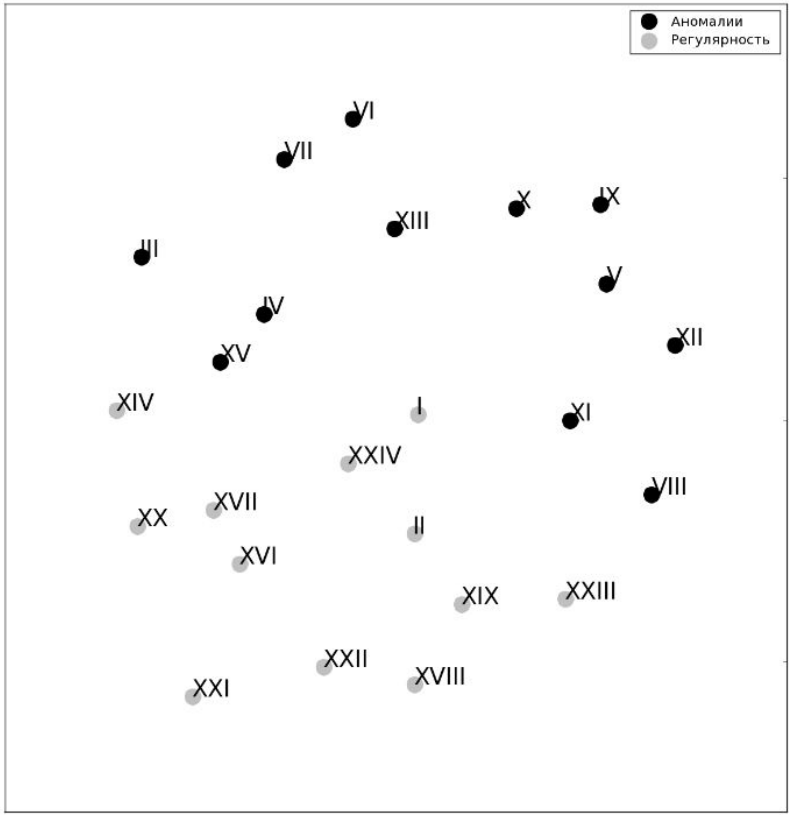
Рис. 2. Соотношение числа встреч и центральности.

Орехов Б. В., Успенский П. Ф., Файнберг В. В. Цифровые подходы к Камер-фурьерскому журналу В. Ф. Ходасевича // Русская литература. — 2018. — № 3. — С. 19–53. DOI: 10.31860/0131-6095-2018-3-19-53

Кластеризация текстов

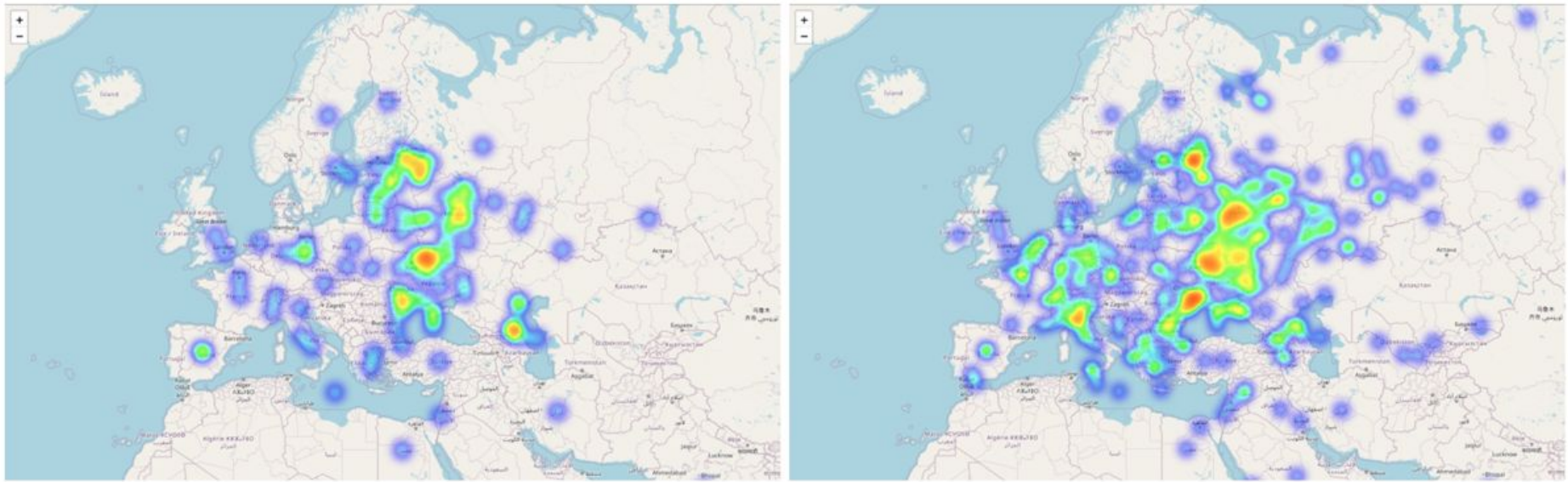


Jamshid J. Tehrani The Phylogeny of Little Red Riding Hood
2013



Орехов Б. В. Почему ошибался Жуковский: о внутритекстовых причинах метрических сбоев в «Одиссее» // М. Л. Гаспарову — стиховеду. In Memoriam. / Сост. Акимова М. В., Тарлинская М. Г. — М.: Языки славянской культуры, 2017. — С. 73–89.

Именованные сущности



Skorinkin, D. & Orekhov, B., (2025) “The Outward Turn. Geocoding the Expansion of Fictional Space in Russian 19th-Century Literature”, Journal of Computational Literary Studies 4(1). doi: https://doi.org/10.48694/jcls.4228

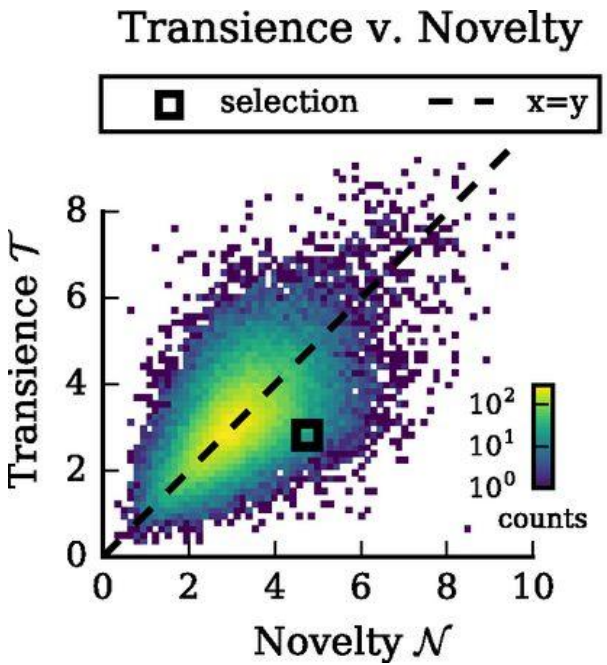
Тематическое моделирование

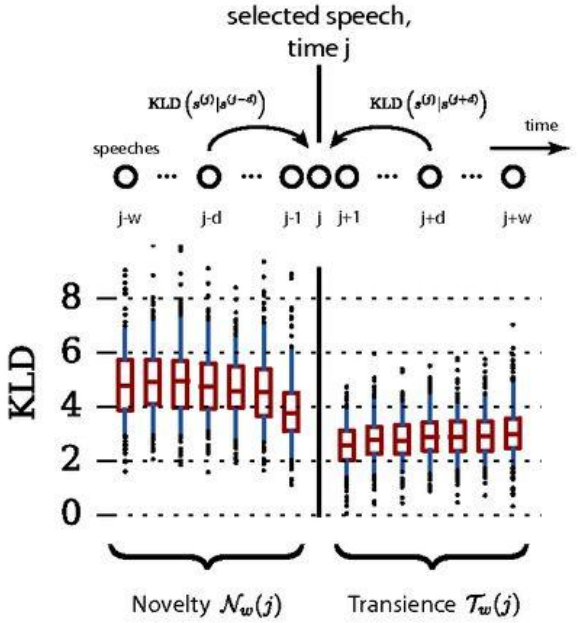
Топики и темы

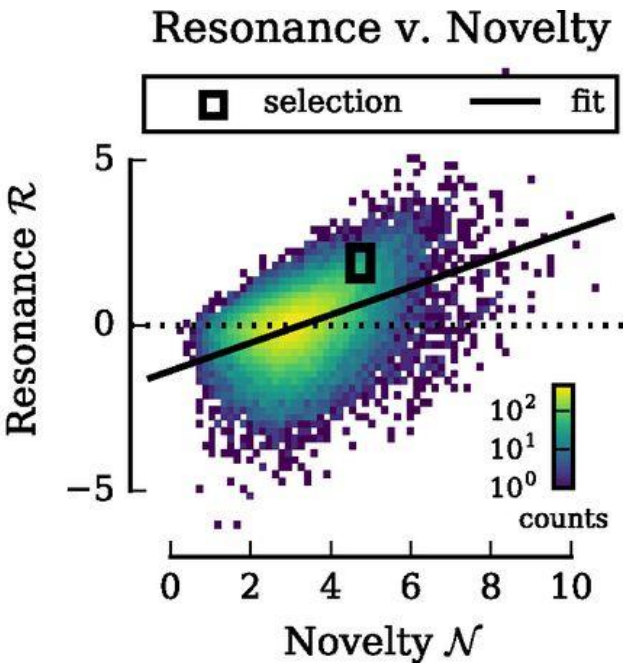
- Используется для задач информационного поиска
- Склеить слова в группы и показывать документы не по словам запроса, а по словам из одной группы
- Исходные данные: совместная встречаемость слов
- Результат: перечень терминов из одного топика

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

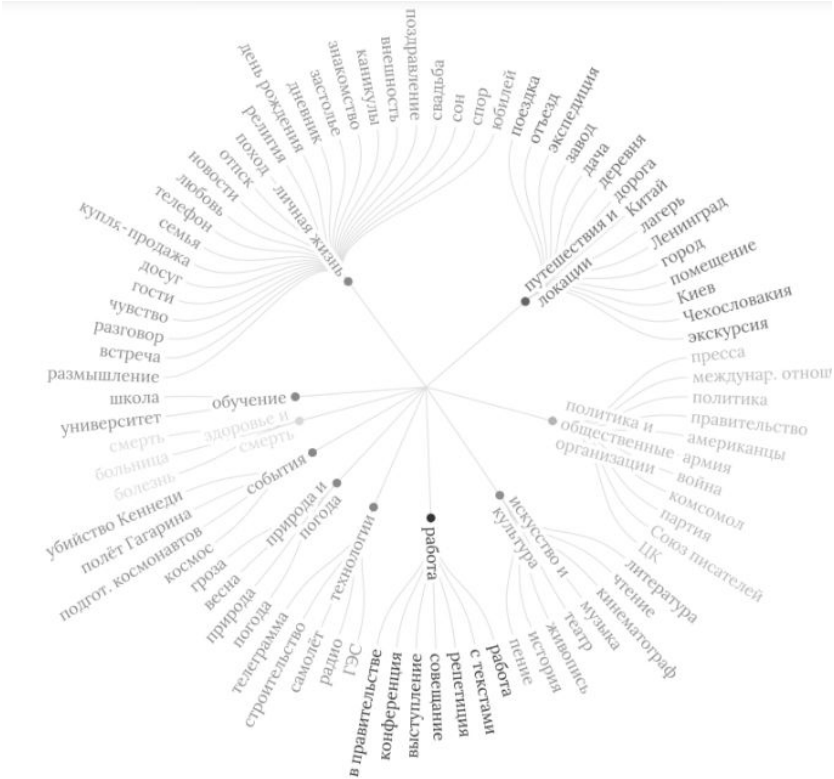
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.







Alexander T. J. Barron, Jenny Huang, Rebecca L. Spang, and Simon DeDeo Individuals, institutions, and innovation in the debates of the French Revolution // PNAS May 1, 2018 115 (18) 4607 – 4612.



Ил. 5. *Радиальное дерево тематических кластеров в корпусе*

Галкина Д. Р., Орехов Б. В., Староверова П. В. О чем пишут в дневниках? Тематическое моделирование на корпусе «Прожито» // Шаги/Steps. — 2026. — 12(3). — С. 278–312. EDN: TUDEOH

Векторные модели

Дистрибутивная семантика

- Слова приобретают значение только в контексте.
- Слова, встречающиеся в похожих контекстах, значат похожее.
- You shall know a word by the company it keeps (J.R. Firth, Papers in Linguistics, 1957)

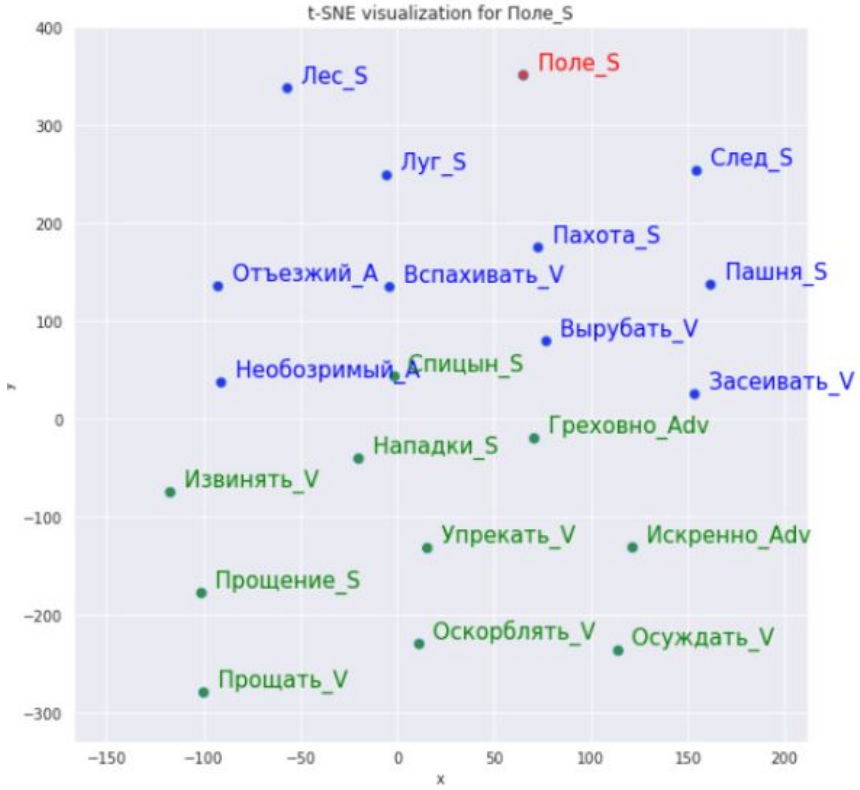
Ср.: Я приду туда через несколько часов vs. Я приду туда через несколько минут



(a) SG, Window=2, $f_{min} = 5$

(b) SG, Window=5, $f_{min} = 5$

Siobh´an Grayson, Maria Mulvany et al. Novel2Vec: Characterising 19th Century Fiction via Word Embeddings // AICS 2016.



Орехов Б. В. Индивидуальная семантика Л. Н. Толстого в свете векторных моделей // Terra Linguistica. — 2023. — Т. 14. — № 4. — С. 119–129. DOI: 10.18721/JHSS.14409

Анализ тональности

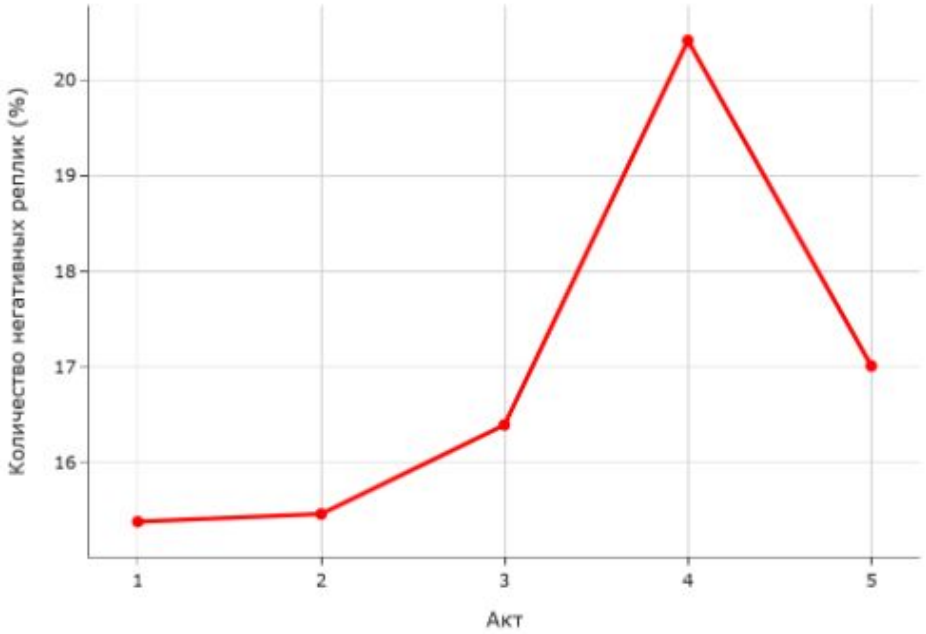


Рис. 6. Негативная динамика пятиактных комедий

Анисимова К. М. Анализ тональности русской драмы XVIII–XX вв. как инструмент моделирования художественной структуры // Цифровые гуманитарные исследования. 2025. № 2. С. 24–47.