

Построение ландшафта области знаний для курса 'Количественные методы в гуманитарных науках: критическое введение' (2024, НИУ ВШЭ)

tinyurl.com/2ydndkzg

Г. А. Мороз
Международная лаборатория языковой конвергенции

13.04.2024

Как все люди Библиотеки, в юности я путешествовал. Это было паломничество в поисках книги, возможно каталога каталогов...

Х. Л. Борхес, "Вавилонская библиотека"

“Информационный гриб”

- Данные вокруг гуманитария?
- Данные вокруг гуманитария...
- Данные вокруг гуманитария!
- Данные вокруг гуманитария:

- Данные вокруг гуманитария?
- Данные вокруг гуманитария...
- Данные вокруг гуманитария!
- Данные вокруг гуманитария: Может быть мы можем попробовать хотя бы обозреть эти самые данные с высоты птичьего полета?

Научных публикаций очень много: желаемое

Google Академия



Стоя на плечах гигантов

Научных публикаций очень много: желаемое

Google Академия



Стоя на плечах гигантов

- “...Мы подобны карликам, усевшимся на плечах великанов; мы видим больше и дальше, чем они, не потому, что обладаем лучшим зрением, и не потому, что выше их, но потому, что они нас подняли и увеличили наш рост собственным величием” высказывание приписывают Бернару Шартрскому, французскому философи XI-XII

Научных публикаций очень много: желаемое

Google Академия



Стоя на плечах гигантов

- “...Мы подобны карликам, усевшимся на плечах великанов; мы видим больше и дальше, чем они, не потому, что обладаем лучшим зрением, и не потому, что выше их, но потому, что они нас подняли и увеличили наш рост собственным величием” высказывание приписывают Бернару Шартрскому, французскому философи XI-XII
- “Today we are privileged to sit side-by-side with the giants on whose shoulders we stand.” Gerald Holton, “On the recent past of physics,” American Journal of Physics, 29 (December, 1961), 805.

Научных публикаций очень много: желаемое

Details



SCIENCE
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE
Volume 134, Issue 3473
Jul 1961

ARTICLE
Impact of Large-Scale Science on the United States
Big science is here to stay, but we have yet to make the hard financial and educational choices it imposes.
[View article page](#)
Alvin M. Weinberg
1961 by the American Association for the Advancement of Science



Throughout history, societies have expressed their aspirations in large-scale, monumental enterprises which, though not necessary for the survival of the societies, have taxed them to their physical and intellectual limits. History often views these monuments as symbolizing the societies. The Pyramids, the Sphinx, and the great temple at Karnak symbolize Egypt; the magnificent cathedrals symbolize the church culture of the Middle Ages; Versailles symbolizes the France of Louis XIV; and so on. The societies were goaded into these extraordinary exertions by their rulers—the pharaoh, the church, the king—who invoked the cultural mystique when this was sufficient, but who also used force when necessary. Sometimes, as with the cathedrals, local

Научных публикаций очень много: желаемое

Details



Science
Volume 134, Issue 3473
Jul 1961

ARTICLE

Impact of Large-Scale Science on the United States

Big science is here to stay, but we have yet to make the hard financial and educational choices it imposes.

[View article page](#)

Alvin M. Weinberg

1961 by the American Association for the Advancement of Science



who also used force when necessary. Sometimes, as with the cathedrals, local pride and a sense of competition with other cities helped launch the project. In many cases the distortion of the economy caused by construction of the big monuments contributed to the civilization's decline.

When history looks at the 20th century, she will see science and technology as its theme; she will find in the monuments of Big Science—the huge rockets, the high-energy accelerators, the high-flux research reactors—symbols of our time just as surely as she finds in Notre Dame a symbol of the Middle Ages. She might even see analogies between our motivations for building these tools of giant science

Научных публикаций очень много: реальность

- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания

Научных публикаций очень много: реальность

- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
 - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит этот эффект

Научных публикаций очень много: реальность

- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
 - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит этот эффект
 - ... люди могут хакнуть и обессмыслить любую метрику

Научных публикаций очень много: реальность

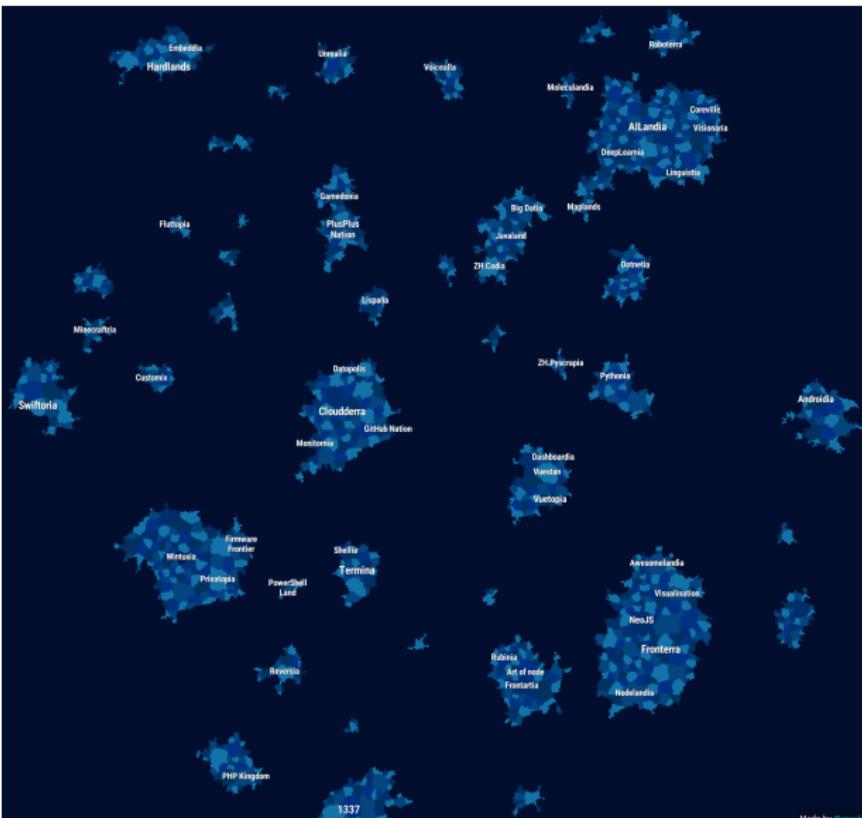
- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
 - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит этот эффект
 - ... люди могут хакнуть и обессмыслить любую метрику
- Исследователи больше любят новые исследования: на материале 726 медицинских статей, содержащих 17 895 научных ссылок, авторы приходят к выводу, что вне зависимости от журнала более 70% цитируемых работ опубликованы не более 10 лет до публикации работы. [Chow et al., 2023]

Научных публикаций очень много: реальность

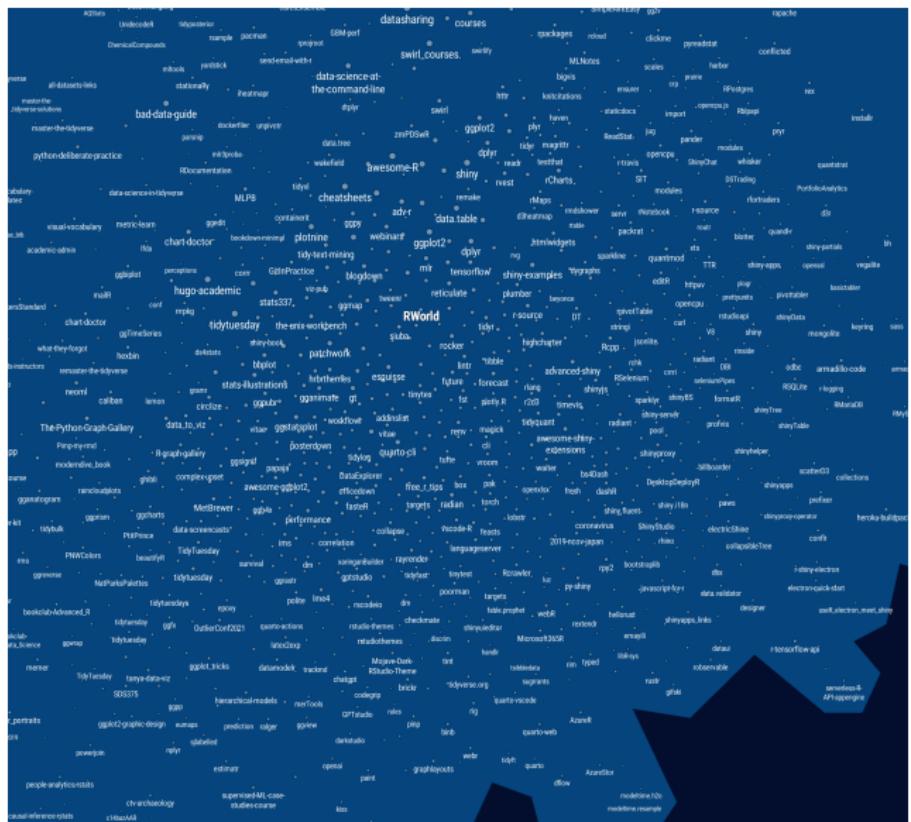
- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
 - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит этот эффект
 - ... люди могут хакнуть и обессмыслиТЬ любую метрику
- Исследователи больше любят новые исследования: на материале 726 медицинских статей, содержащих 17 895 научных ссылок, авторы приходят к выводу, что вне зависимости от журнала более 70% цитируемых работ опубликованы не более 10 лет до публикации работы. [Chow et al., 2023]
- Даже цифра может подгнить: авторы обнаружили значительную долю “мертвых” URL статей, которые упоминаются при цитировании в публикациях в медицине. [Klein et al., 2014]

Ландшафты

Карта репозиториев гитхаба (Андрей Кашча)



Карта репозиториев гитхаба (Андрей Кашча)



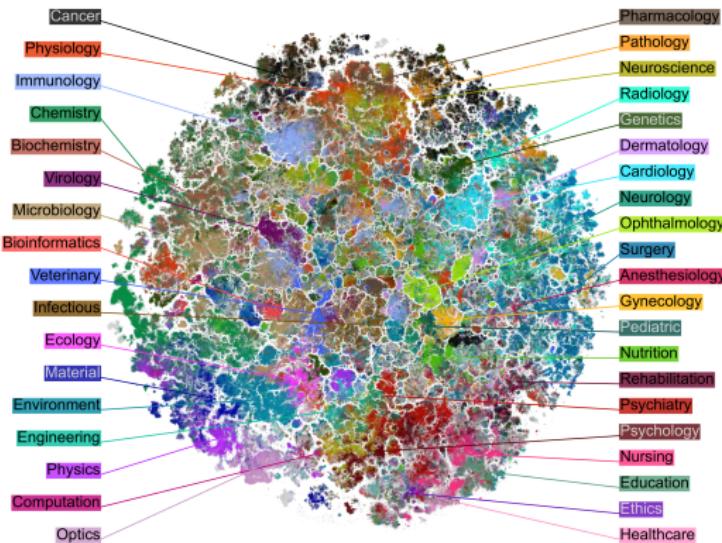
[Gonzalez-Marquez et al., 2023]

The number of publications in biomedicine and life sciences has rapidly grown over the last decades, with over 1.5 million papers now published every year. This makes it difficult to keep track of new scientific works and to have an overview of the evolution of the field as a whole. Here we present a 2D atlas of the entire corpus of biomedical literature, and argue that it provides a unique and useful overview of the life sciences research. <...>

<https://static.nomic.ai/pubmed.html> (интерактивная версия)

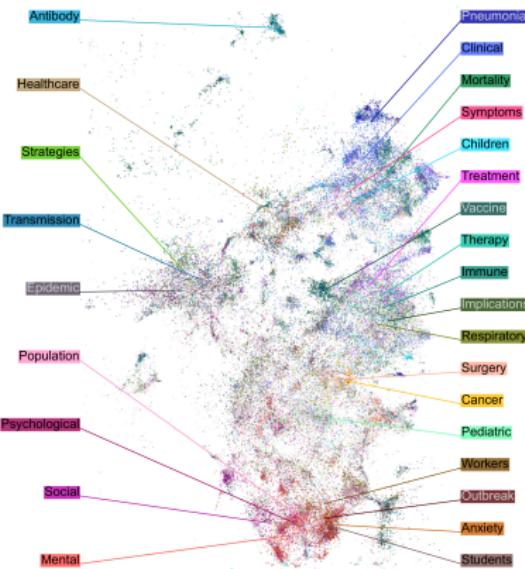
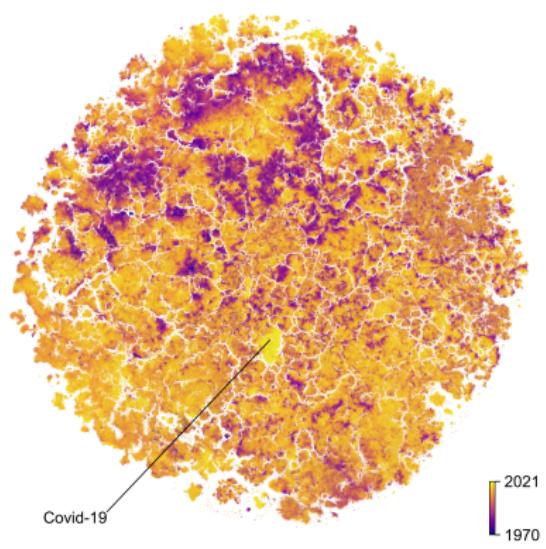
Это препринт!

[Gonzalez-Marquez et al., 2023]



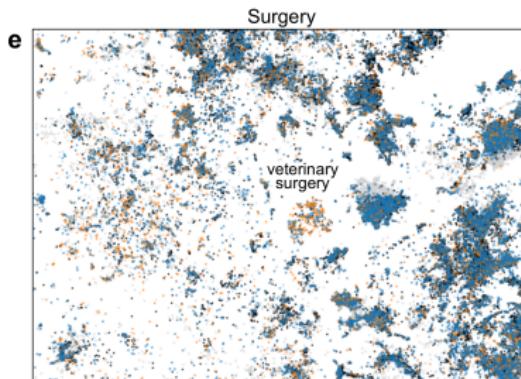
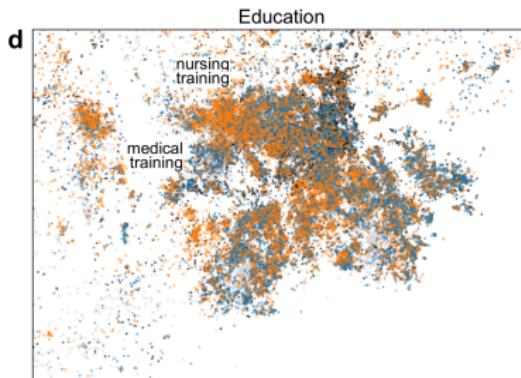
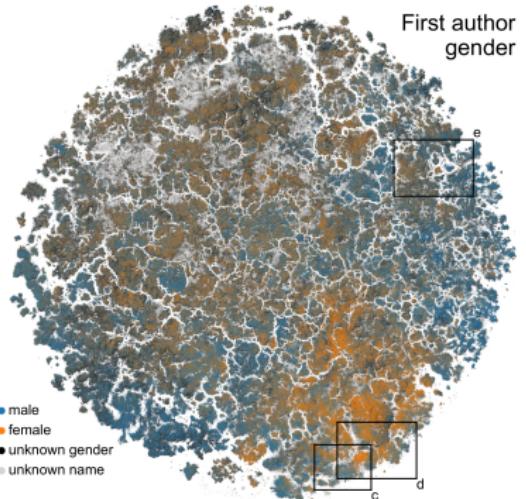
2D эмбеддинги на основе 21 миллиона аннотаций, которые были трансформированы в 768-мерное векторное пространство при помощи PubMedBERT [Gu et al., 2021], а дальше сплюснутая в 2D при помощи t-SNE [Van der Maaten and Hinton, 2008]. Цвета основаны на названиях журналов.

[Gonzalez-Marquez et al., 2023]



Регион карты, посвященный Covid-19. Цвета приписаны на основе названий работ. Кроме того здесь есть около 15% работ не посвященных короновирусу.

[Gonzalez-Marquez et al., 2023]



Статьи раскрашены по полу первого автора.

Другие проекты Nomic

- map of Wikipedia
- map of Twitter
- другие <https://atlas.nomic.ai/discover>

Похожее

Библиометрические исследования?

Библиометрия — дисциплина, возникшая в конце XIX века, в рамках которой можно встретить разные применения математических методов к исследованию научных работ. Наиболее известные применения:

- графы соавторства
- библиографические ссылки
- ключевые слова
- измерение качества журналов
- и др.

Distant Reading?

Дальнее чтение [Moretti, 2013] — это не какой-то один метод, а целое семейство методов анализа литературных текстов и их структуры, а также подразумевающий некоторый осмыслиенный с точки зрения литературоведения исследовательский вопрос.

"Информационный гриб"
○○○○○

Ландшафты
○○○○○○○○

Похожее
○○○

Техническое
●○○○○○○

Исследование лингвистики
○○○○○○○○○○○○○○○○

Ограничения метода
○○○○○○○

References

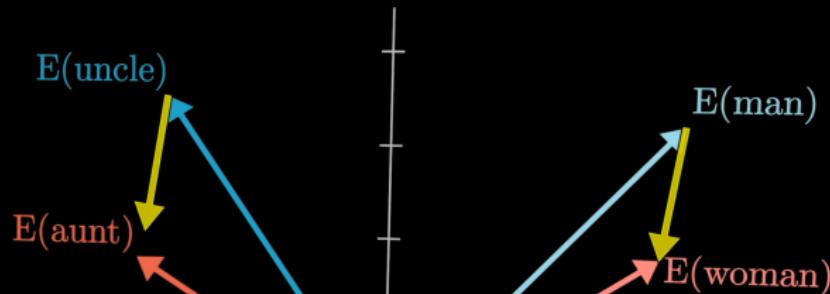
Техническое

Эмбеддинги

- Архитектурам машинного обучения любой сложности при работе с языковыми данными нужно уметь преобразовывать слова (на самом деле некоторые кусочки письменных слов) в наборы чисел, которые обычно называют **вектором**.
- Числа для вектора каждого конкретного слова обычно получают на основе контекстов, в которых оно появляется в обучающем корпусе.
- Слова с похожим значением будут направлены в одну сторону. Сравнивать их следует по углу между векторами.
- В работах [Mikolov et al., 2013a,b] от Google была представлена модель word2vec, архитектура нейросети для создания векторных моделей.
- Совсем недавно вышли видео 3Blue1Brown, в которых это обсуждается подробнее:
 - But what is a GPT? Visual intro to transformers
 - Attention in transformers, visually explained

Эмбеддинги

$$E(aunt) - E(uncle) \approx E(woman) - E(man)$$



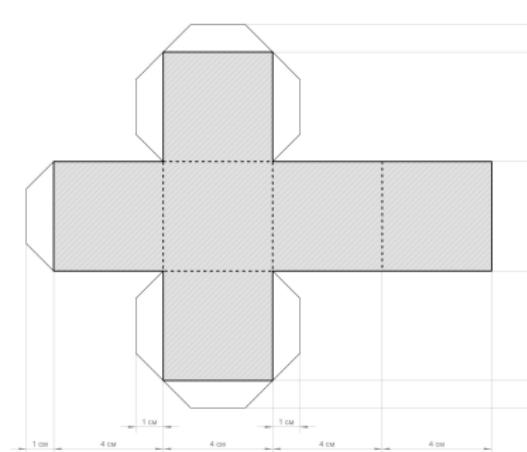
Взято из видео 3Blue1Brown.

doc2vec

- Чтобы анализировать тексты [Le and Mikolov, 2014] предложили разбивать их на абзацы и конкатенировать векторы, которые входят в абзац, а потом использовать их для кластеризации текстов.
- Если применять эту логику к предложениям, то это позволяет не терять информацию о месте слова.

Уменьшение размерности

- Эмбеддинги — многомерные вектора чисел, например, в GPT-3 50 тысяч токенов закодировано при помощи векторов длиной 12 тысяч. Смотреть на это пространство глазами нельзя, но можно попробовать уменьшить размерность.



Уменьшение размерности

- Эмбеддинги — многомерные вектора чисел, например, в GPT-3 50 тысяч токенов закодировано при помощи векторов длиной 12 тысяч. Смотреть на это пространство глазами нельзя, но можно попробовать уменьшить размерность.
- Популярные алгоритмы:
 - Principal Component Analysis (PCA)
 - Multidimensional Scaling (MDS)
 - Linear discriminant analysis (LDA)
 - Uniform Manifold Approximation and Projection (UMAP)
 - t-distributed Stochastic Neighbor Embedding (t-SNE)

Кластеризация

- Полученные группы в пространстве часто можно выделить автоматически. Для этого используют методы кластеризации, чаще всего Hierarchical Density-based spatial clustering of applications with noise (HDBSCAN) [Ester et al., 1996, Campello et al., 2013]

Исследование лингвистики

Команда

- руководитель
 - Г. Мороз
- студенты
 - А. Агроскина (б)
 - Т. Дедов (б)
 - А. Орехов (м)
 - К. Сидоров (м)
 - А. Степанова (б)

План исследования

- выбрать список журналов для анализа
- извлечь аннотации для всех работ из выбранных журналов
- использовать векторизатор и метод уменьшения размерностей для преобразования пространства аннотаций в 2D
- исследовать, насколько релевантно для лингвистики получившееся пространство
- выявить и исследовать возможные междисциплинарные стыки

Списки журналов

Мы использовали несколько источников журналов

- Тэг филология, лингвистика, медиакоммуникации в вышкинском списке журналов

Списки журналов НИУ ВШЭ / HSE Journal Lists

Список А / List A	Список В / List B	Список С / List C	Список D / List D
627	43	1941	36

Название ISSN Сп... Категория

ANC: AUGMENTATIVE AND ALTERNATIVE COMMU...	0743-9618; 1417-...	А	ФИЛОЛОГИЯ, МЕДИЦИНА И ЗДРАВООХРАНЕНИЕ; ФИЛО...
ACROSS LANGUAGES AND CULTURES	1585-1623; 1586-...	А	ФИЛОЛОГИЯ, ЛИНГВИСТИКА И МЕДИАКОММУНИКАЦИИ...
ACTA BOREALIA	0806-3831; 1503-...	А	ИСКУССТВО И ГУМАНИТАРНЫЕ НАУКИ; ИСТОРИЯ, АРХ...

- Тэг 6162 Languages в списке журналов из ресурса [Finish Publication Forum](#)

Results 1 - 20 / 1245 First Previous Next Last

LevelTitle

AALITRA REVIEW
2 ACROSS LANGUAGES AND CULTURES
1 ACROSS THE DISCIPLINES
1 ACTA ACUSTICA

Списки журналов

После соединения списков журналов мы по своему усмотрению разметили их по некоторым категориям (теги: linguistics (358), interdisciplinary (433), language_learning (69) и другие).

		Helsenki level		
HSE level		a	b	c
a		42	37	15
b		1	3	10
c		0	24	124
d		0	1	4

Списки журналов

После соединения списков журналов мы по своему усмотрению разметили их по некоторым категориям (теги: linguistics (358), interdisciplinary (433), language_learning (69) и другие).

		Helsenki level		
HSE level		a	b	c
a		42	37	15
b		1	3	10
c		0	24	124
d		0	1	4

Разметка "лингвистичности" журналов — огромная и слабо автоматизируемая работа, которая требует экспертизы в самых разных областях лингвистики.

Сбор аннотаций лингвистических исследований

- Мы планировали написать краулер, который бы собирал статьи из желаемых журналов...

Сбор аннотаций лингвистических исследований

- Мы планировали написать краулер, который бы собирал статьи из желаемых журналов...
- но мы обнаружили базу данных Crossref и соответствующий пакет для R `rcrossref` [[Chamberlain et al., 2022](#)]...

Сбор аннотаций лингвистических исследований

- Мы планировали написать краулер, который бы собирал статьи из желаемых журналов...
- но мы обнаружили базу данных Crossref и соответствующий пакет для R `rcrossref` [[Chamberlain et al., 2022](#)]...
- а потом мы обнаружили базу данных OpenAlex и соответствующий пакет для R `openalexR` [[Aria and Le, 2023](#)]

Чистка аннотаций

- заметки редактора

Чистка аннотаций

- заметки редактора
- некрологи и поздравления

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей
- аннотации на отличном от английского языках

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей
- аннотации на отличном от английского языках
- аннотации на нескольких языках

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей
- аннотации на отличном от английского языках
- аннотации на нескольких языках
- сообщения об отсутствии аннотации

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей
- аннотации на отличном от английского языках
- аннотации на нескольких языках
- сообщения об отсутствии аннотации
- acknowledgments вместо аннотации

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей
- аннотации на отличном от английского языках
- аннотации на нескольких языках
- сообщения об отсутствии аннотации
- acknowledgments вместо аннотации
- библиографическое описание книги (в случаях рецензии)

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей
- аннотации на отличном от английского языках
- аннотации на нескольких языках
- сообщения об отсутствии аннотации
- acknowledgments вместо аннотации
- библиографическое описание книги (в случаях рецензии)
- начало статьи вместо аннотации (характерно для старых статей)

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей
- аннотации на отличном от английского языках
- аннотации на нескольких языках
- сообщения об отсутствии аннотации
- acknowledgments вместо аннотации
- библиографическое описание книги (в случаях рецензии)
- начало статьи вместо аннотации (характерно для старых статей)
- ошибки распознавания

Чистка аннотаций

- заметки редактора
- некрологи и поздравления
- описания конференций
- списки содержания книг
- списки содержания выпусков журнала
- аннотации отмененных (retracted) статей
- аннотации на отличном от английского языках
- аннотации на нескольких языках
- сообщения об отсутствии аннотации
- acknowledgments вместо аннотации
- библиографическое описание книги (в случаях рецензии)
- начало статьи вместо аннотации (характерно для старых статей)
- ошибки распознавания
- слишком короткие/длинные аннотации

Примеры проблемных аннотаций

SpringerLink

<https://doi.org/10.1007/BF02743731>

Search [Login](#)

[Home](#) > [Russian Linguistics](#) > Article

Articles | Published: June 1990

ъзыкъ БытА / ъзыкъИ ДУХОВ НОИ кУЛЬТУРы

[Russian Linguistics](#) 14, 129–146 (1990) | [Cite this article](#)

18 Accesses | 3 Citations | [Metrics](#)

This is a preview of subscription content, [access via your institution](#).

лИтЕРАтУРА

БАРт, Р.: 1978, 'лИнгВИстИкА тЕкстA'.НОВОЕ В яЗАРУБ ЕжНОИ лИнгВИстИкЕ. Вып. VIII. лИнгВИстИкА тЕкст A, МОСКВА, 442–449.

..

Access via your institution

Access options

Buy article PDF

39,95 €

Price includes VAT (Russian Federation)

Instant access to the full article PDF.

[Rent this article via DeepDyve.](#)

Примеры проблемных аннотаций

LANGUAGE OF THE TOBACCO MARKET

ROBERT J. FITZPATRICK

Louisville, Kentucky

LISTEN to the chant of the tobacco auctioneer.' 'Fo'teen-a-lee-di-leen-a-
lee-di-leen — — — qwa-qwa-qwa-qwa-aw-aw — — — ha-ha-ha-ha-ha
— — — three-di-lee-di-lee — — fifteen — American.' How familiar is this
chant to the listeners of a well-known radio program. Yet how many of
them could tell if they heard the same jargon at a real tobacco auction
that the bid on a pile of tobacco had been opened at fourteen dollars a
hundred pounds, that the buyers had raised the bid to \$14.25, to \$14.50,
to \$14.75, and that the tobacco had finally been sold at \$15.00?

<https://doi.org/10.2307/486818>

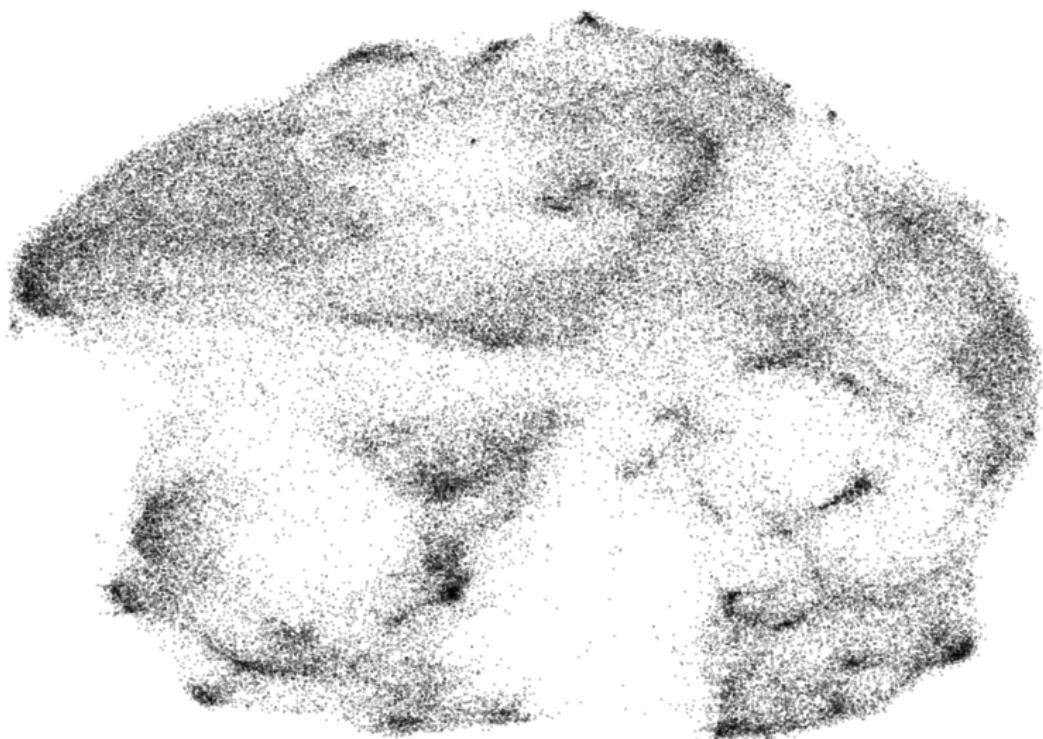
Структура данных: 79545 строчек, 24 колонок

- **id:** <https://openalex.org/W3o4o6uu73o>
- **doi:** <https://doi.org/10.1075/fol.18056.dob>
- **author:** Nina Dobrushina
- **title:** Negation in complement clauses of fear-verbs
- **publication_year:** 2021
- **journal:** Functions of Language
- **issn_l:** 0929-998X
- **first_page:** 121
- **last_page:** 152
- **volume:** 28
- **issue:** 2
- **is_retracted:** FALSE
- **cited_by_count:** 1
- **abstract:** Complement clauses of verbs of fear often contain expletive negation, which is negative marking without negative meaning. <...>
- **concepts:** Negation; Complement (music); Linguistics; Verb; Meaning (existential); Psychology; Mathematics;

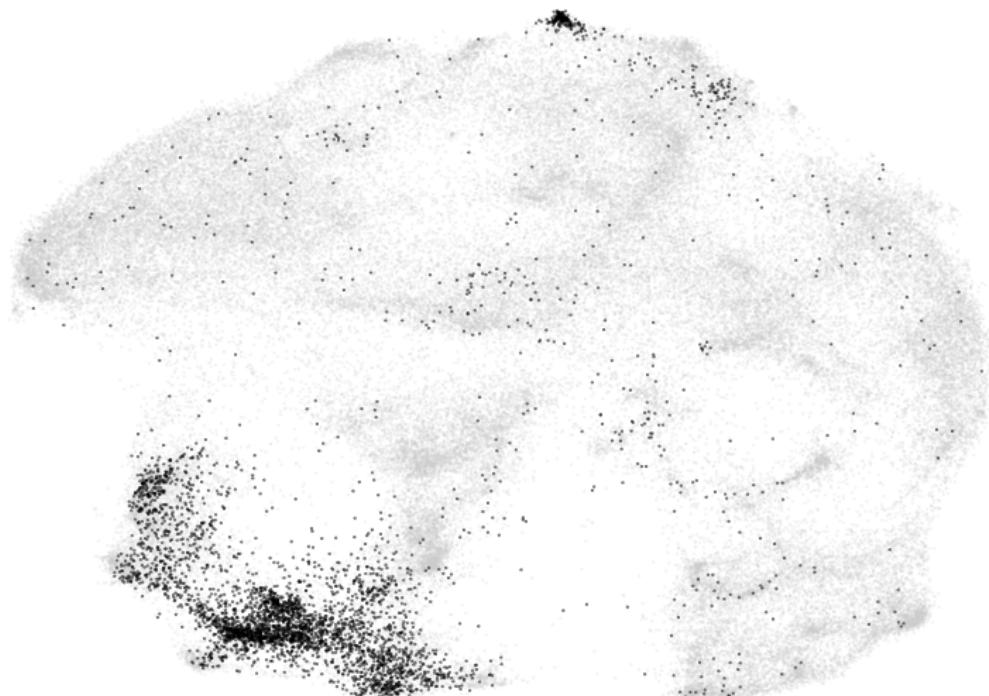
Векторное представление слов

- Мы использовали векторизатор doc2vec [[Le and Mikolov, 2014](#), [Wijffels, 2021](#)] (смотрели GloVe [[Pennington et al., 2014](#)], BERT [[Devlin et al., 2018](#)] и RoBERTa [[Liu et al., 2019](#)])
- Полученное 50-мерное пространство мы сократили до 2D при помощи UMAP [[McInnes et al., 2018](#)]

Ландшафт лингвистических исследований



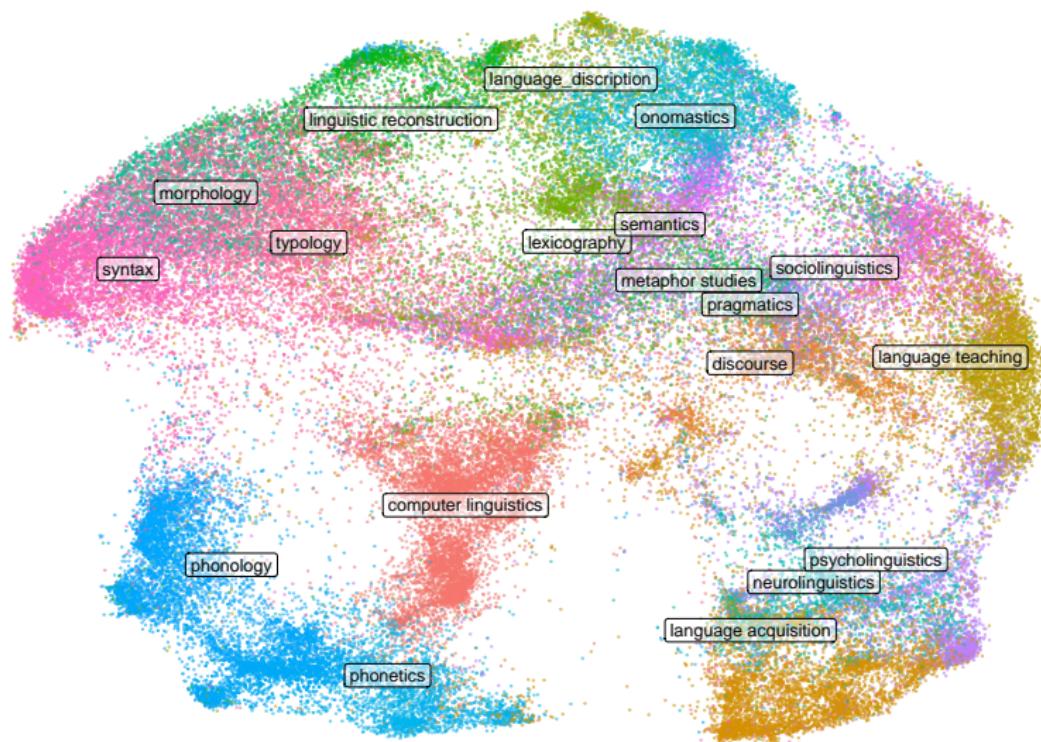
Аннотации журналов, в названиях которых содержится *phon*



Аннотации журналов, в названиях которых содержится *psych*



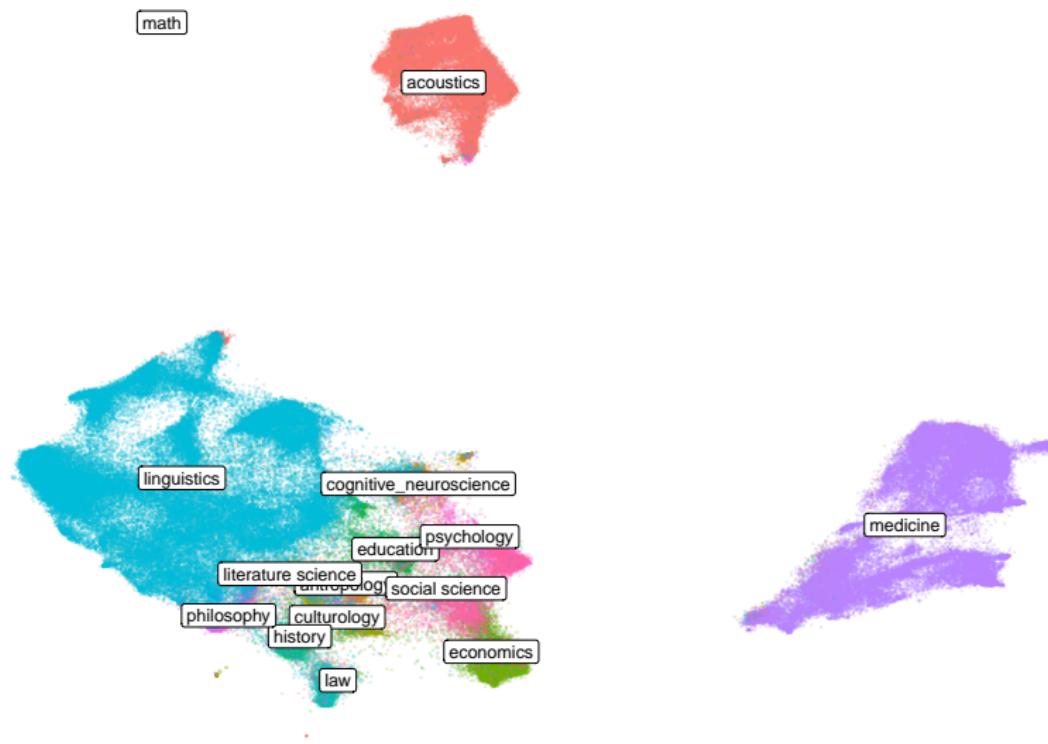
Наша полуавтоматическая разметка аннотаций



А что если добавить других дисциплин?

field	journal	n
acoustics	Applied Acoustics	7511
acoustics	Journal of Sound and Vibration	22071
anthropology	Annual Review of Anthropology	915
anthropology	Current Anthropology	1719
cognitive_neuroscience	Biological Psychiatry: Cognitive Neuroscience and Neuroimaging	818
cognitive_neuroscience	Trends in Cognitive Sciences	2495
culturology	European Journal of Cultural Studies	1014
culturology	The Journal of Peasant Studies	1422
economics	Journal of Political Economy	2654
economics	The American Economic Review	6405
education	Educational Research Review	1313
education	Educational Researcher	1579
education	Review of Educational Research	1953
history	Annales. Histoire, Sciences Sociales	330
history	History	578
history	The Historical Journal	2297
law	American Journal of International Law	2758
law	Berkeley Journal of International Law	106
law	European Journal of International Law	1339
literature science	American Journal of Philology	683
literature science	Poetics	1406
math	NA	1750
medicine	The Lancet	50736
philosophy	American Philosophical Quarterly	350
philosophy	Journal of the History of Philosophy	2062
psychology	Annual Review of Psychology	871
psychology	Journal of Applied Psychology	3523
psychology	Psychological Bulletin	1894
social science	Administrative Science Quarterly	1539
social science	American Sociological Review	3932

А что если добавить других дисциплин?

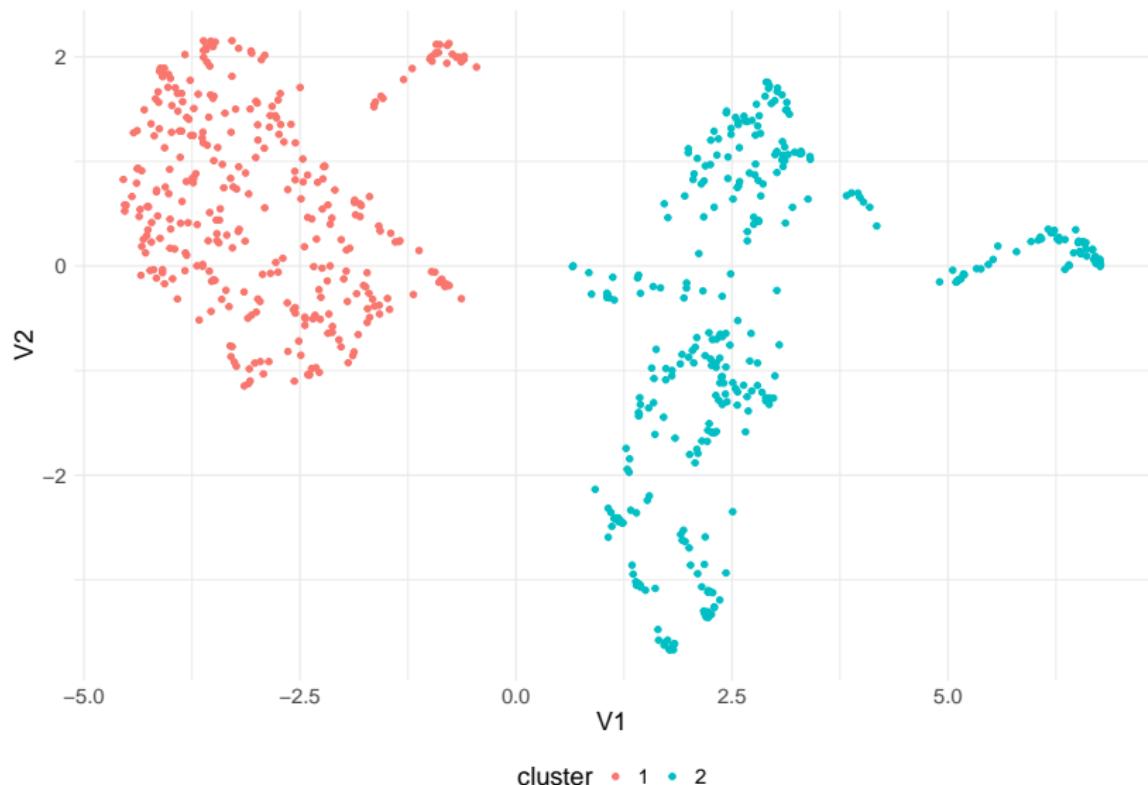


Ограничения метода

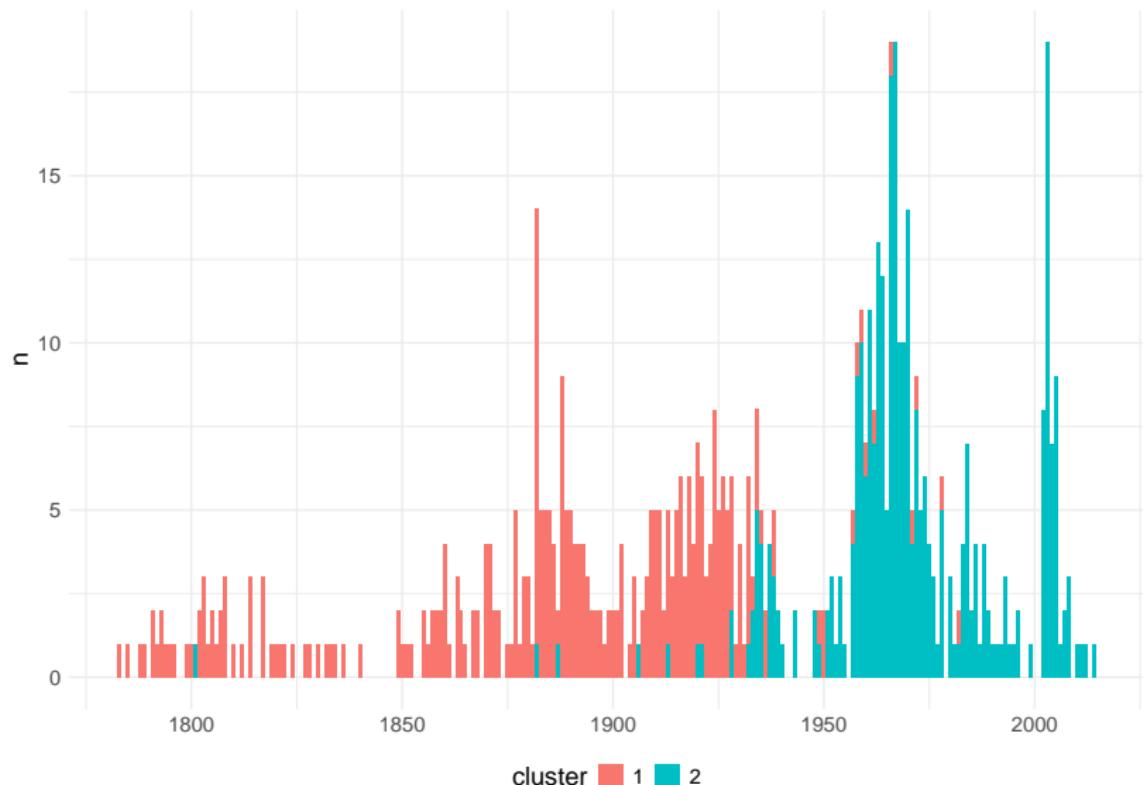
Ограничения метода:

- количество текстов должно быть большое
 - мы безуспешно пробовали процедуру на уровне отдельных российских журналов
- нужно понимать контекст текстов

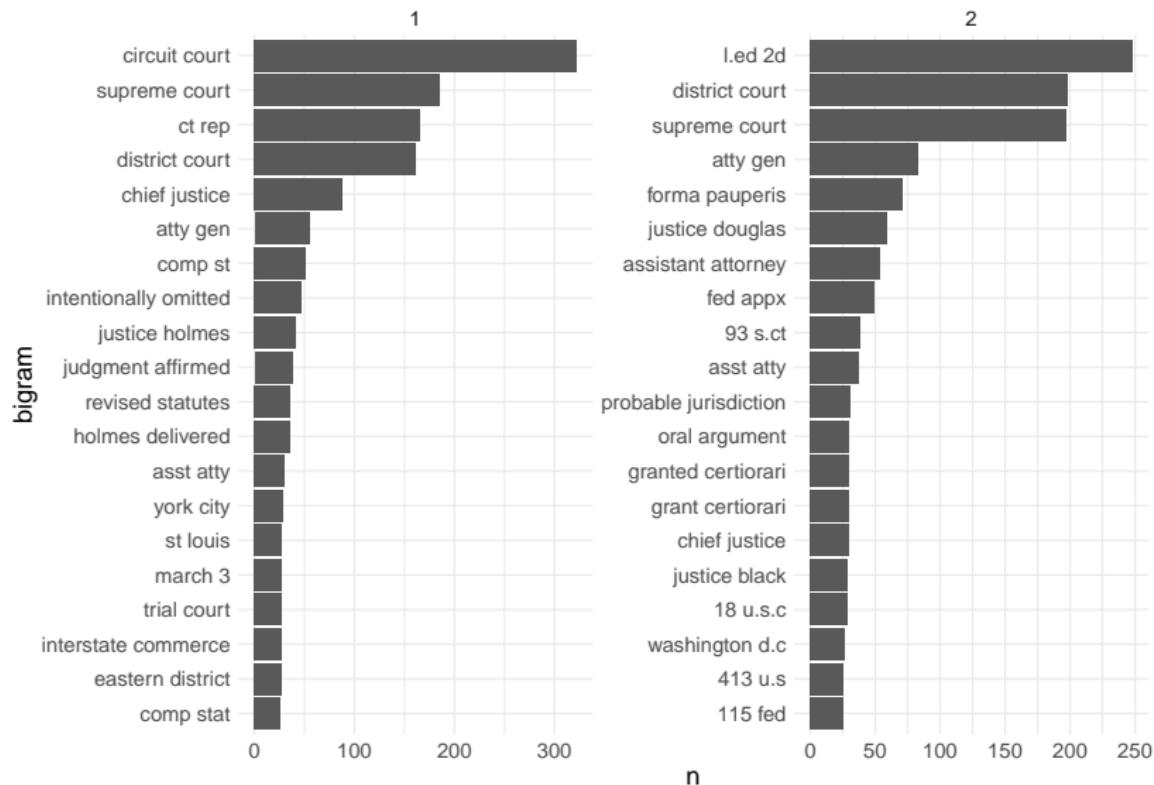
Ограничения метода: 648 мнений верховного суда США



Ограничения метода: 648 мнений верховного суда США



Ограничения метода: 648 мнений верховного суда США



Ограничения метода: 648 мнений верховного суда США (1)

No. 769. See 353 U.S. 989, 77 S.Ct. 1281. Mr. Thomas D. McBride, Atty. Gen. of Pennsylvania, and Lois G. Forer, Deputy Atty. Gen., Philadelphia, for Commonwealth of Pennsylvania. Messrs. Abraham L. Freedman and David Berger, Philadelphia, for City of Philadelphia and others. Messrs. William T. Coleman, Jr., Raymond Pace Alexander, Philadelphia, and Louis Pollak, for appellants Foust and others. Mr. Owen B. Rhoads, Philadelphia, for appellee. PER CURIAM. The motion to dismiss the appeal for want of jurisdiction is granted. 28 U.S.C. § 1257(2), 28 U.S.C.A. § 1257(2). Treating the papers whereon the appeal was taken as a petition for writ of certiorari, 28 U.S.C. § 2103, 28 U.S.C.A. § 2103, the petition is granted. 28 U.S.C. § 1257(3), 28 U.S.C.A. § 1257(3). Stephen Girard, by a will probated in 1831, left a fund in trust for the erection, maintenance, and operation of a college. The will provided that the college was to admit as many poor white male orphans, between the ages of six and ten years, as the said income shall be adequate to maintain. The will named as trustee the City of Philadelphia. The provisions of the will were carried out by the State and City and the college was opened in 1848. Since 1869, by virtue of an act of the Pennsylvania Legislature, the trust has been administered and the college operated by the Board of Directors of City Trusts of the City of Philadelphia. Pa.Laws 1869, No. 1258, p. 1276; Purdons Pa.Stat.Ann., 1957, Tit. 53, § 16365. In February 1954, the petitioners Foust and Felder applied for admission to the college. They met all qualifications except that they were Negroes. For this reason the Board refused to admit them. They petitioned the Orphans Court of Philadelphia County for an order directing the Board to admit them, alleging that their exclusion because of race violated the Fourteenth Amendment to the Constitution. The State of Pennsylvania and the City of Philadelphia joined in the suit also contending the Boards action violated the Fourteenth Amendment. The Orphans Court rejected the constitutional contention and refused to order the applicants admission. *In re Girards Estate*, 4 Pa.Dist. & Co.R.2d 671 (Orph.Ct.Philadelphia). This was affirmed by the Pennsylvania Supreme Court. 386 Pa. 548, 127 A.2d 287. The Board which operates Girard College is an agency of the State of Pennsylvania. Therefore, even though the Board was acting as a trustee, its refusal to admit Foust and Felder to the college because they were Negroes was discrimination by the State. Such discrimination is forbidden by the Fourteenth Amendment. *Brown v. Board of Education*, 347 U.S. 483, 74 S.Ct. 686, 98 L.Ed. 873. Accordingly, the judgment of the Supreme Court of Pennsylvania is reversed and the cause is remanded for further proceedings not inconsistent with this opinion. Reversed and remanded with directions.

Ограничения метода: 648 мнений верховного суда США (2)

No. 330. Mr. Nash Rockwood, of New York City, for appellants. Mr. John Caldwell Myers, Asst. Dist. Atty., of New York City, for appellees. Mr. Chief Justice TAFT delivered the opinion of the Court. This appeal is from an order of the District Court for the Southern District of New York discharging a rule nisi and refusing an injunction. On January 14th a petition in involuntary bankruptcy was filed against Elmore D. Dier and others, partners, as E. D. Dier & Co. Two days after the filing of the petition, Mandred Ehrich was appointed receiver of the estate of the alleged bankrupts, and they and their servants were directed to turn over all their property, assets, account books and records, and were restrained from suing out of any other court any process to impound or take possession of them. This order was complied with and the receiver took possession of the books and papers of the alleged bankrupts and of the firm. On February 16th, Dier informed the court that the district attorney of New York City had applied to the receiver for the production of these books and papers before the grand jury, and asked for the rule nisi against the receiver and the district attorney, and upon a hearing thereof an injunction to prevent the use of such books and papers against him before the grand jury, on the ground that they would incriminate him and that his right to refuse to testify against himself under the Fourth and Fifth Amendments would thus be violated by the process of the federal District Court. Judge Learned Hand, sitting in bankruptcy, discharged the rule and refused to enjoin the proposed use of the books. Judge Hands action was based on the ruling of this court in *Johnson v. United States*, 228 U. S. 457, 33 Sup. Ct. 572, 57 L. Ed. 919, 47 L. R. A. (N. S.) 263. He quoted the language used in the Johnson Case: A party is privileged from producing the evidence, but not from its production. He alluded to the circumstance that in the Johnson Case there were both title and possession in the trustee, whereas in this case the books and papers were in the hands of the receiver, who has no title; but that, he said, made no difference. We agree with this view, and hold that the right of the alleged bankrupt to protest against the use of his books and papers relating to his business as evidence against him ceases as soon as his possession and control over them pass from him by the order directing their delivery into the hands of the receiver and into the custody of the court. This change of possession and control is for the purpose of properly carrying on the investigation into the affairs of the alleged bankrupt and the preservation of his assets pending such investigation, the adjudication of bankruptcy vel non, and, if bankruptcy is adjudged, the proper distribution of the estate. It may be that the allegation of bankruptcy will not be sustained, and in that case the alleged bankrupt will be entitled to a return of his property, including his books and papers, and, when they are returned, he may refuse to produce them and stand on his constitutional rights. But while they are, in the due course of the bankruptcy proceedings, taken out of his possession and control, his immunity from producing them, secured him under the Fourth and Fifth Amendments, does not inure to his protection. He has lost any right to object to their use as evidence because, not for purpose of evidence, but in the due investigation of his alleged bankruptcy and the preservation of his estate pending such investigation, the control and possession of his books and papers relating to his business were lawfully taken from him. It is pressed upon us that the bankrupt may prevent the use of such books and papers taken over by a receiver in the bankruptcy proceedings for evidence in a criminal case in the state court by resisting surrender and protesting against their use for such a purpose at the time the receiver took possession. But we think the alleged bankrupt has no such right. We so held in *Matter of Fuller*, 262 U. S. 91, 43 Sup. Ct. 496, 67 L. Ed. ——, decided April 30, 1923, in which it was sought to attach conditions of this kind to the turning over of the books and papers of a bankrupt to the trustee in bankruptcy. We are of opinion that the same principle must

*Мне известен дикий край, где библиотекари отказались от суеверной
и напрасной привычки искать в книгах смысл, считая, что это все
равно что искать его в снах или в беспорядочных линиях руки...*

Х. Л. Борхес, “Вавилонская библиотека”

Список литературы I

Massimo Aria and Trang Le. *openalexR: Getting Bibliographic Records from 'OpenAlex' Database Using 'DSL' API*, 2023. URL <https://CRAN.R-project.org/package=openalexR>. R package version 1.0.2.9.

Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the association for information science and technology*, 66(11):2215–2222, 2015.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

Список литературы II

Scott Chamberlain, Hao Zhu, Najko Jahn, Carl Boettiger, and Karthik Ram.

rcrossref: Client for Various 'CrossRef' APIs', 2022. URL

<https://CRAN.R-project.org/package=rcrossref>. R package version 1.2.0.

Natalie LY Chow, Natalie Tateishi, Alexa Goldhar, Rabia Zaheer, Donald A Redelmeier, Amy H Cheung, Ayal Schaffer, and Mark Sinyor. Does knowledge have a half-life? an observational study analyzing the use of older citations in medical and scientific publications. *BMJ open*, 13(5):e072374, 2023.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Список литературы III

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231, 1996.

Rita Gonzalez-Marquez, Luca Schmidt, Benjamin M Schmidt, Philipp Berens, and Dmitry Kobak. The landscape of biomedical research. *bioRxiv*, 2023. doi: <https://doi.org/10.1101/2023.04.10.536208>.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Список литературы IV

Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. Scholarly context not found: one in five articles suffers from reference rot. *PloS one*, 9(12): e115253, 2014.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Список литературы V

Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013b.

Franco Moretti. *Distant reading*. Verso Books, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Список литературы VI

Derek J. de Solla Price. *Little science, big science*. Columbia University Press, 1963.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

Jan Wijffels. *doc2vec: Distributed Representations of Sentences, Documents and Topics*, 2021. URL <https://CRAN.R-project.org/package=doc2vec>. R package version 0.2.0.