

[Gonzalez-Marquez et al., 2024]

для журнального клуба EBM_BASE

Г. А. Мороз

Международная лаборатория языковой конвергенции НИУ ВШЭ

11.05.2024

Обо мне

Лингвист

- полевой исследователь (30 экспедиций, почти все на Кавказ)
- фонетист, фонолог, квантитативный лингвист, занимаюсь лингвистической географией
- преподаю статистику и R
- написал несколько лингвистических пакетов для R
 - `lingtypology`
 - `phonfieldwork`
 - `lingglosses`

Лингвист

- полевой исследователь (30 экспедиций, почти все на Кавказ)
- фонетист, фонолог, квантитативный лингвист, занимаюсь лингвистической географией
- преподаю статистику и R
- написал несколько лингвистических пакетов для R
 - `lingtypology`
 - `phonfieldwork`
 - `lingglosses`
- я не занимаюсь обучением и исследованием больших языковых моделей

Лингвист

- полевой исследователь (30 экспедиций, почти все на Кавказ)
- фонетист, фонолог, квантитативный лингвист, занимаюсь лингвистической географией
- преподаю статистику и R
- написал несколько лингвистических пакетов для R
 - `lingtypology`
 - `phonfieldwork`
 - `lingglosses`
- я не занимаюсь обучением и исследованием больших языковых моделей
- 99% лингвистов не занимается придумыванием, какого рода кофе и как нужно говорить или писать

Rationale

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 философу 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 философу 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: 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

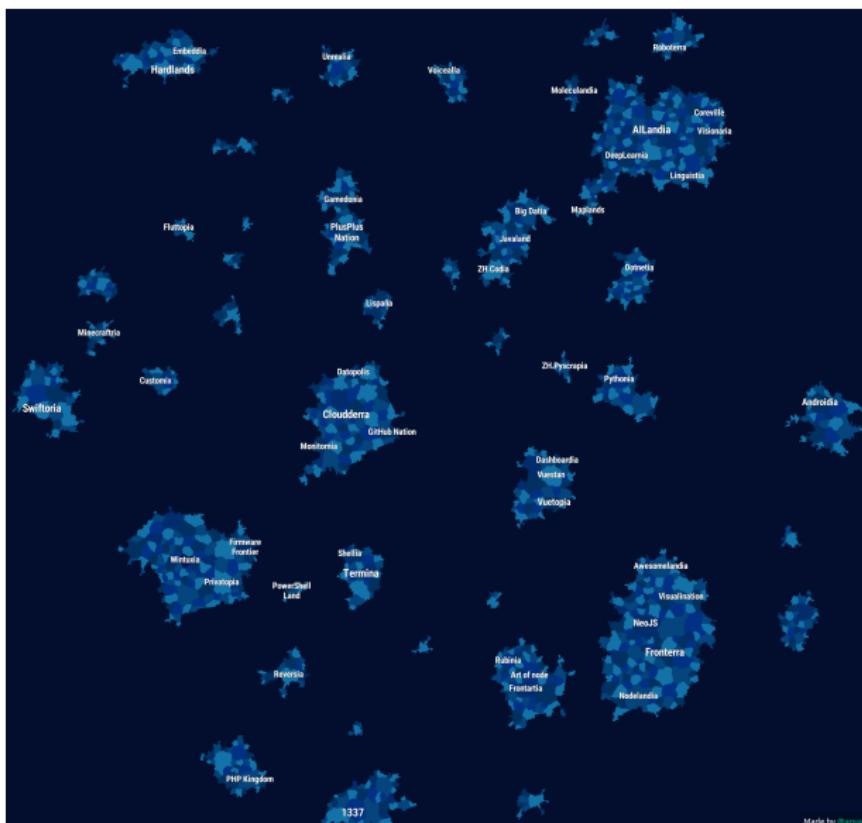
- Динамика сохраняется: [Price, 1963, Bornmann and Mutz, 2015]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
 - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит этот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

Научных публикаций очень много: реальность

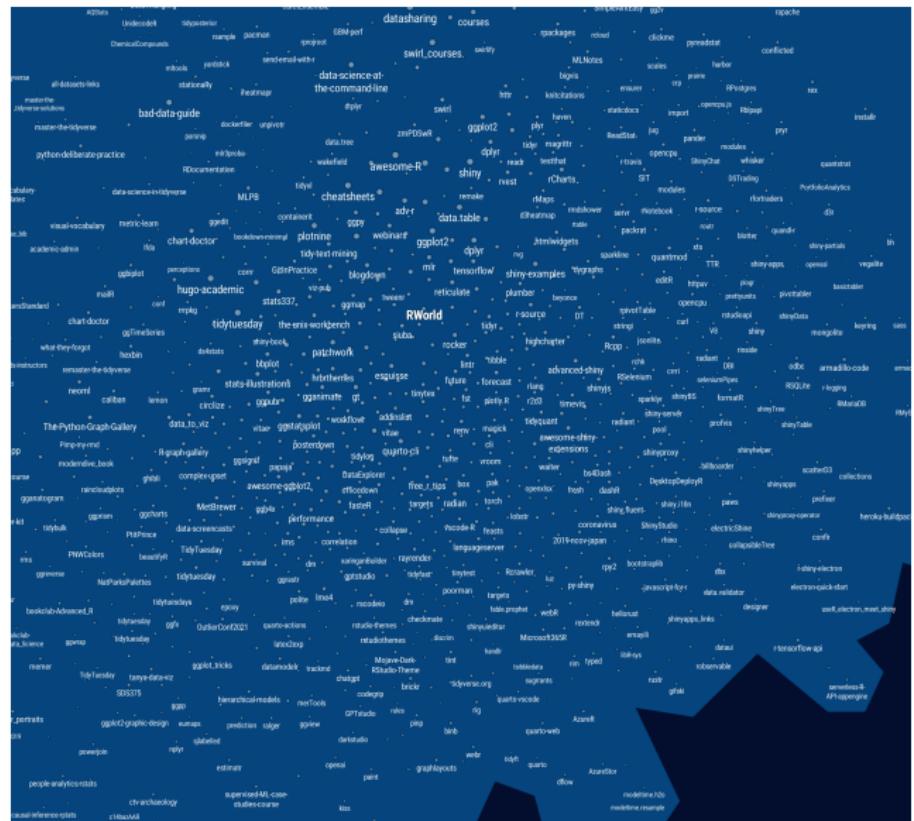
- Динамика сохраняется: [[Price, 1963](#), [Bornmann and Mutz, 2015](#)]
- Очень сложно разобраться в какой-либо области знания
- Количество цитирований (или другие библиометрические меры) могли бы помочь, но ...
 - ... люди все чаще цитируют, не читая и эра больших языковых моделей скорее всего увеличит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

Ландшафты

Карта репозиториев гитхаба (Андрей Кашча)



Карта репозиториев гитхаба (Андрей Кашча)



[Gonzalez-Marquez et al., 2024]

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)

Проекты Nomic

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

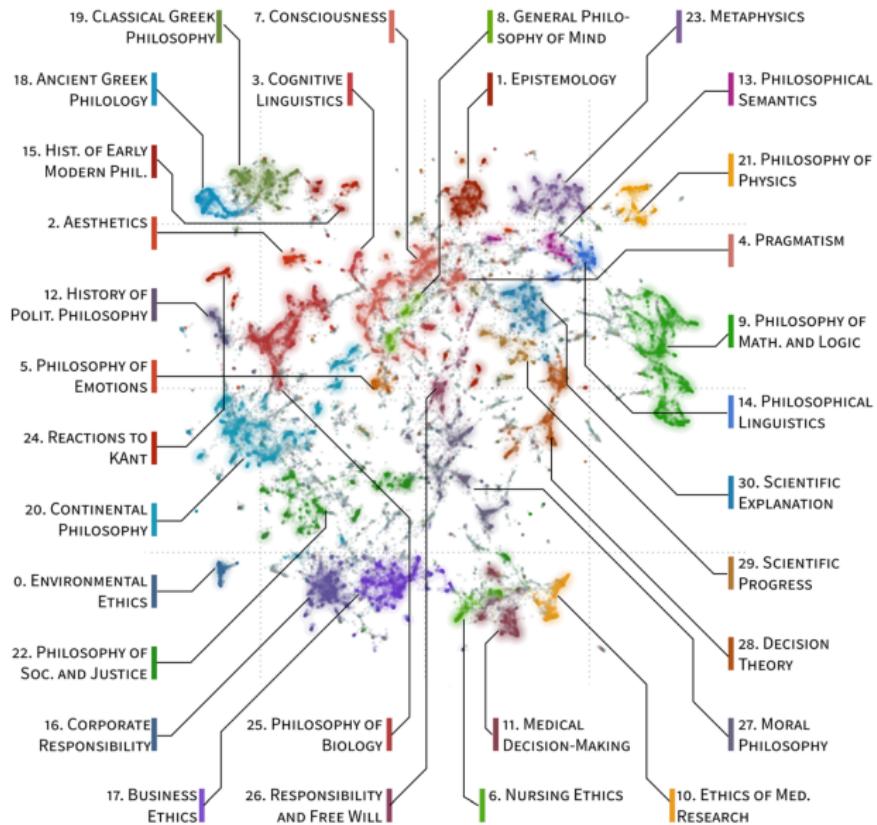
Похожее

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.

[Noichl, 2021]



[Noichl, 2021]

Работа очень похожа целями и результатом на работу [Gonzalez-Marquez et al., 2024], но основано на списках литературы: если авторы ссылаются на сходных авторов, значит они, вероятно, будут близки в полученном пространстве.

Обсуждение [Gonzalez-Marquez et al., 2024]

Журнал

Статья опубликована в журнале Patterns:

Patterns is a premium open access journal from Cell Press, publishing ground-breaking original research across the full breadth of data science. We're all about sharing data science solutions to problems that cross domain boundaries.

Patterns reaches a broad, global audience of computer scientists, researchers in data intensive domains, data stewards, and policy makers. We adhere to the FAIR Principles to make sure that the data, software, workflows, algorithms, and other research outputs we publish are findable, accessible, interoperable, and reusable.

Patterns is the home for data scientists and researchers in data-intensive fields in both academia and industry. The journal shares data science solutions across the spectrum of disciplines, including computational, physical, life, and social sciences, and the humanities.

Авторы статьи

- Rita González-Márquez^{1,2}
- Luca Schmidt^{1,2}
- Benjamin M. Schmidt³
- Philipp Berens^{1,2}
- Dmitry Kobak^{1,5}

- 1) Hertie Institute for AI in Brain Health, University of Tübingen, Germany
- 2) Tübingen AI Center, Tübingen, Germany
- 3) Nomic AI, New York, New York, USA
- 4) IWR, Heidelberg University, Heidelberg , Germany
- 5) Lead contact

Авторы статьи

- Rita González-Márquez^{1,2}
- Luca Schmidt^{1,2}
- Benjamin M. Schmidt³
- Philipp Berens^{1,2}
- Dmitry Kobak^{1,5}

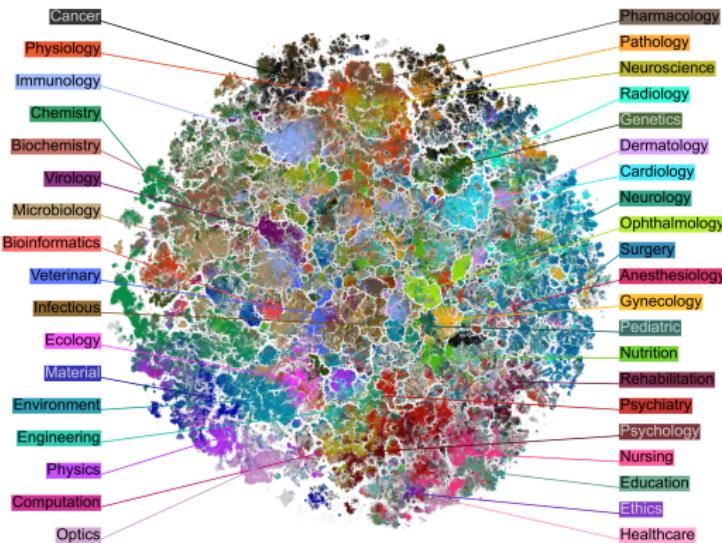
- 1) Hertie Institute for AI in Brain Health, University of Tübingen, Germany
- 2) Tübingen AI Center, Tübingen, Germany
- 3) Nomic AI, New York, New York, USA
- 4) IWR, Heidelberg University, Heidelberg , Germany
- 5) Lead contact

Авторы статьи

- Rita González-Márquez^{1,2}
- Luca Schmidt^{1,2}
- Benjamin M. Schmidt³
- Philipp Berens^{1,2}
- Dmitry Kobak^{1,5}

	R.G.-M.	L. S.	B. M. S.	P. B.	D. K.
Design the study	+				+
Analysis and Figures	+				
Pilot experiments with LMs		+			
Interactive website			+		
Initial draft	+				
Edited the paper	+			+	+
Discussed the results	+			+	+
Study supervision				+	+

Что было сделано



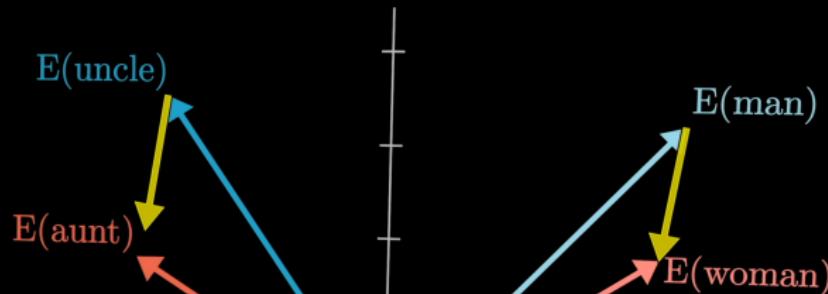
2D пространство на основе 21 миллиона аннотаций (snapshot 2021 года), которые были трансформированы в 768-мерное векторное пространство при помощи PubMedBERT [Gu et al., 2021], а дальше сплюснуты в 2D при помощи t-SNE [Van der Maaten and Hinton, 2008]. Цвета основаны на названиях журналов.

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

$$E(\text{aunt}) - E(\text{uncle}) \approx E(\text{woman}) - E(\text{man})$$



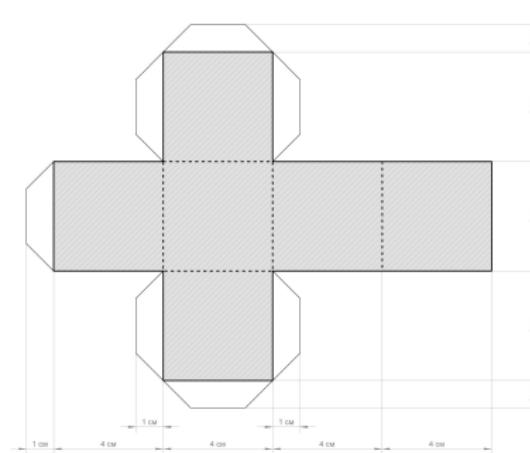
Взято из видео 3Blue1Brown.

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)

Важно еще обратить внимание [на видео](#) Дмитрия Кобака, посвященное уменьшению размерностей (начиная с 56 минуты обсуждается сегодняшняя статья).

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]

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]
- При этом авторы использовали кластеризацию несколько по-другому:
 - они сначала автоматически аннотировали аннотации на основе названия журнала (точнее 34.4% аннотаций)
 - а дальше при помощи алгоритма k -NN смотрят насколько разделимы они в многомерном пространстве.

?? Авторы написали, что взяли весь PubMed

The landscape of biomedical research



?? Авторы написали, что взяли весь PubMed

The landscape of biomedical research



MY NCBI FILTERS

RESULTS BY YEAR

RESULTS: 11,281 results

CAR-T cell therapy: current lim...
Sterner RC, Sterner RM.
Blood Cancer J. 2021 Apr 6;11(4):69. doi:
PMID: 33824268 **Free PMC article**
Chimeric antigen receptor (CAR)-T cell...
addition, the host and tumor microenv...
cell function. Furthermore, a compl ...

1987 2024

?? Frontiers in Immunology (April, May 2024)

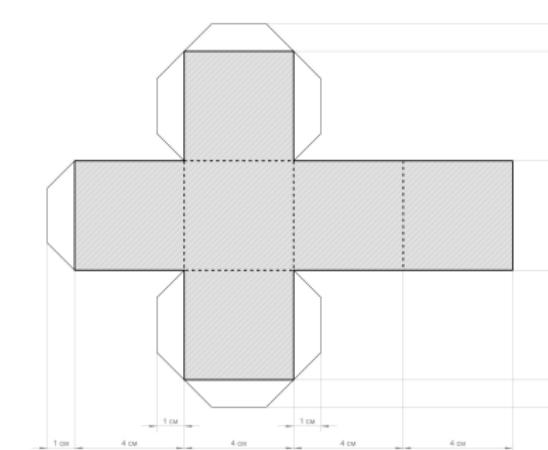
- Identification and Validation of Biomarkers in Membranous Nephropathy and Pan-Cancer Analysis
- Material basis and molecular mechanisms of Chaihuang Qingyi Huoxue Granule in the treatment of acute pancreatitis based on network pharmacology and molecular docking-based strategy
- MiR-146a alleviates inflammatory bowel disease in mice through systematic regulation of multiple genetic networks
- Screening mitochondria-related biomarkers in skin and plasma of atopic dermatitis patients by bioinformatics analysis and machine learning
- Independent organelle and organelle—organelle interactions: essential mechanisms for malignant gynecological cancer cell survival
- Single-cell transcriptome reveals highly complement activated microglia cells in association with pediatric tuberculous meningitis
- ...

?? Метрика качества

Авторы в качестве метрики качества используют точность (accuracy) k -NN кластеризации 34.4% аннотаций.

Однако разумным, как мне кажется, должна быть локальная и глобальная структура, которую бы оценивали биологи и врачи, т. е. насколько осмысленным является соседство тех или иных областей.
Однако сложно представить

Case study: COVID



Список литературы I

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.

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.

Список литературы II

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. *Patterns*, 5, 2024. doi: <https://doi.org/10.1016/j.patter.2024.100968>.

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 III

- Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.
- 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.
- Maximilian Noichl. Modeling the structure of recent philosophy. *Synthese*, 198(6):5089–5100, 2021.

Список литературы IV

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.