

Построение ландшафта лингвистики: анализ аннотаций лингвистических статей

Г. Мороз, А. Агроскина, А. Алексеева, Т. Дедов, А. Орехов, К.
Сидоров, А. Степанова

16 июня 2023

План доклада

Введение

Наш проект

Сбор данных

Описательная статистика

Ландшафт лингвистических исследований

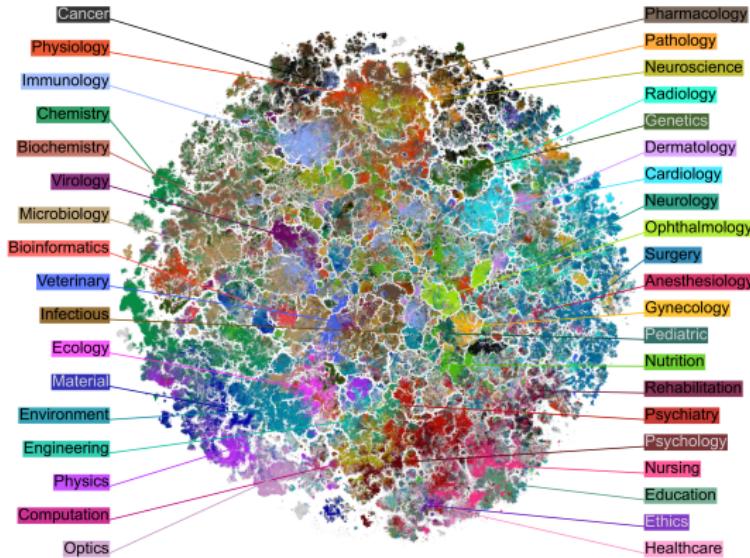
Заключение

(Gonzalez-Marquez et al. 2023) ландшафт биомедицинских исследований

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. <...>

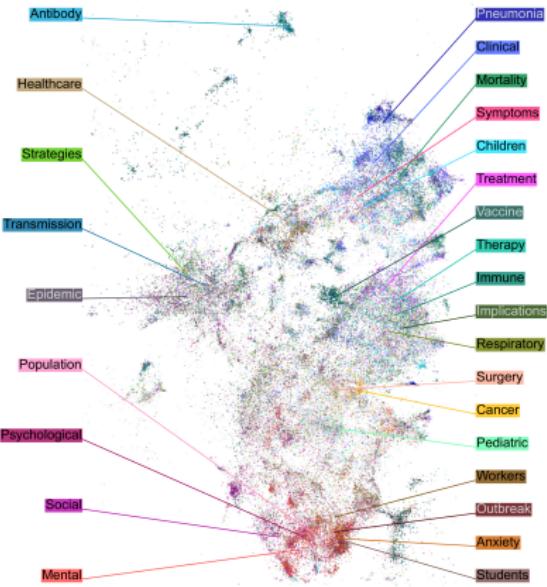
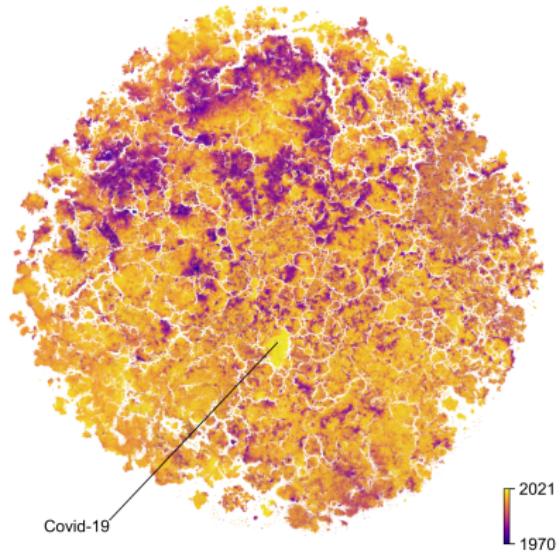
<https://static.nomic.ai/pubmed.html> (интерактивная версия)

Это препринт!

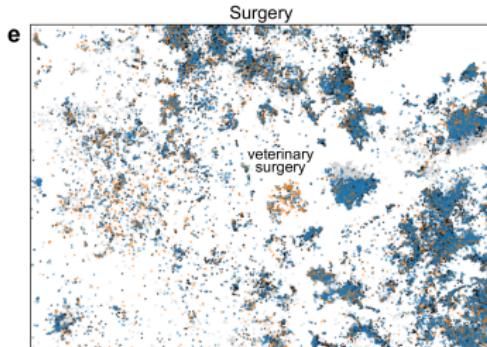
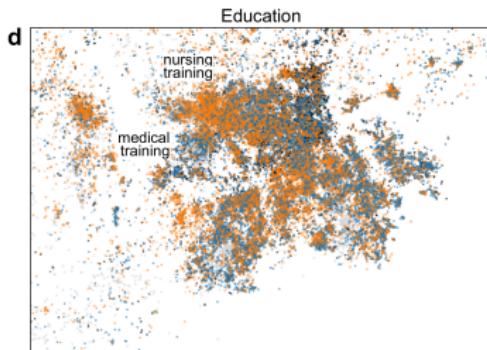
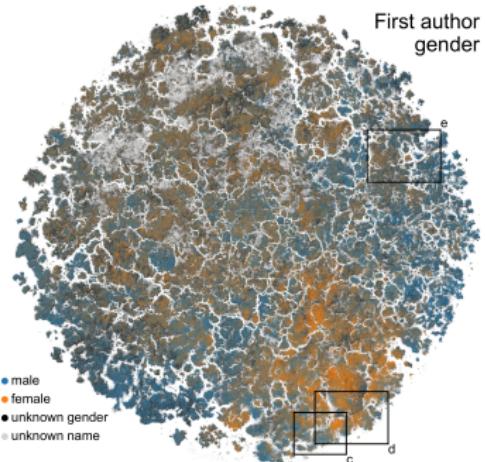


2D эмбеддинги на основе 21 миллиона аннотаций, которые были трансформированы в 768-мерное векторное пространство при помощи PubMedBERT (Gu et al. 2021), а дальше сплюснутая в 2D при помощи t-SNE (Van der Maaten and Hinton 2008). Цвета основаны на названиях журналов.

(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.



Статьи раскрашены по полу первого автора.

План доклада

Введение

Наш проект

Сбор данных

Описательная статистика

Ландшафт лингвистических исследований

Заключение

План

- 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

План доклада

Введение

Наш проект

Сбор данных

Описательная статистика

Ландшафт лингвистических исследований

Заключение

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 Б / List B	Список С / List C	Список D / List D
627	43	1941	36

Название ISSN Сп... Категория

AAC: AUGMENTATIVE AND ALTERNATIVE COMMU...	0744-9618; 1477-...	А	БИОЛОГИЯ, МЕДИЦИНА И ЗДРАВООХРАНЕНИЕ; ФИЛО...
ACROSS LANGUAGES AND CULTURES	1585-1923; 1568-...	А	ФИЛОЛОГИЯ, ЛИНГВИСТИКА И МЕДИАКОММУНИКАЦИИ...
ACTA BOREALIA	0806-3831; 2503-...	А	ИСКУССТВО И ГУМАНИТАРНЫЕ НАУКИ; ИСТОРИЯ, АРХИ...

- 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 (тэги: linguistics (418), interdisciplinary (414), language_learning (43) и другие).

Отдельный подпроект: литературоведение (тэги: literary_studies (432), interdisciplinary (503), philology (12) и другие).

После разметки 418 (15%) журналов из 2750 получили тэг linguistics.

HSE level	Helsenki level				
	a	b	c	d	NA
a	46	44	19	1	1
b	1	3	12	0	3
c	0	25	140	1	94
d	0	1	5	0	4
NA	1	17	0	0	0

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](#))

План доклада

Введение

Наш проект

Сбор данных

Описательная статистика

Ландшафт лингвистических исследований

Заключение

Структура данных: 350576 строчек, 19 колонок

- **id:** <https://openalex.org/W3040611730>
- **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; Computer science; Philosophy; Biochemistry; Chemistry; Complementation; Psychotherapist; Gene; Phenotype
- **tags_level:** 2; 5; 1; 2; 2; O; O; O; O; O; 1; O; 4; 1; 2; 3

Чистка данных

Нам пришлось достаточно существенно почистить данные:

- аннотации не на английском языке
- аннотации, отмененных (*retracted*) статей
- размеченные как аннотации информационные сообщения журналов
- мусор, который отмечен как аннотация

... так что осталось 103696 строк.

План доклада

Введение

Наш проект

Сбор данных

Описательная статистика

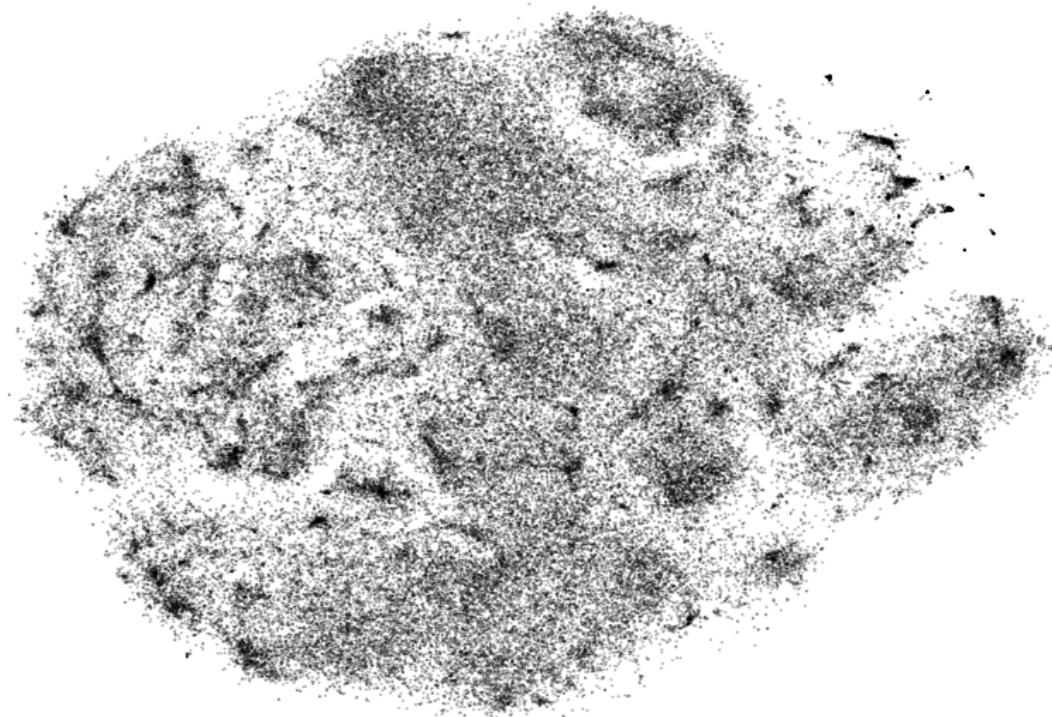
Ландшафт лингвистических исследований

Заключение

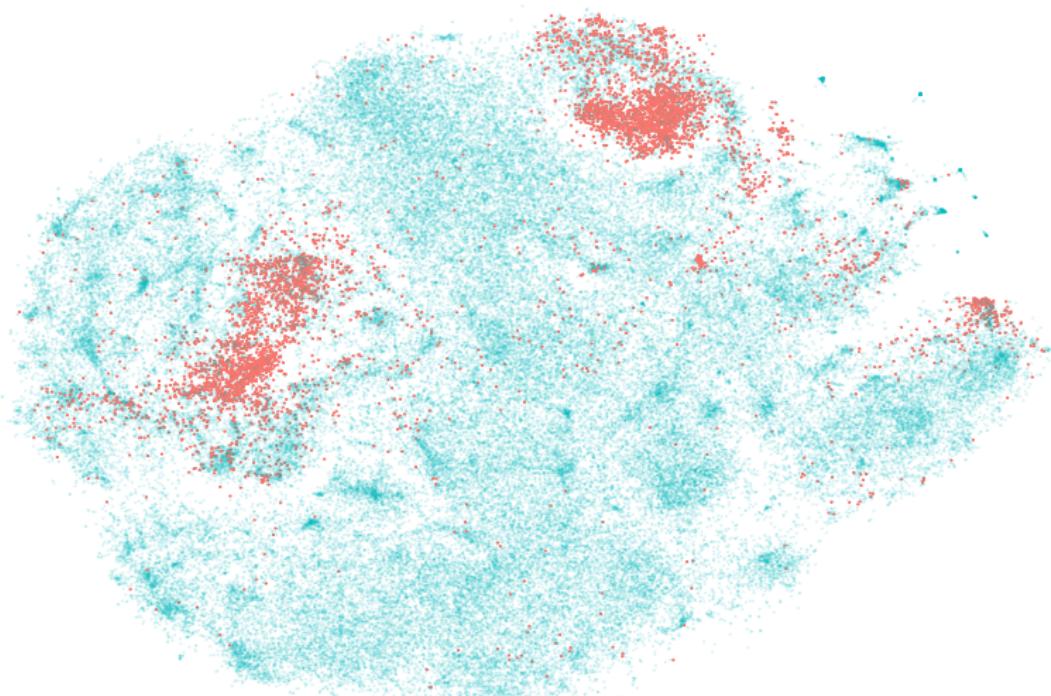
Векторное представление слов

- мы использовали векторизатор doc2vec (Le and Mikolov 2014; Wijffels 2021) (смотрели GloVe (Pennington et al. 2014), думаем в сторону BERT (Devlin et al. 2018) и RoBERTa (Liu et al. 2019))
- полученное 50-мерное пространство мы схлопывали при помощи t-SNE (Van der Maaten and Hinton 2008)

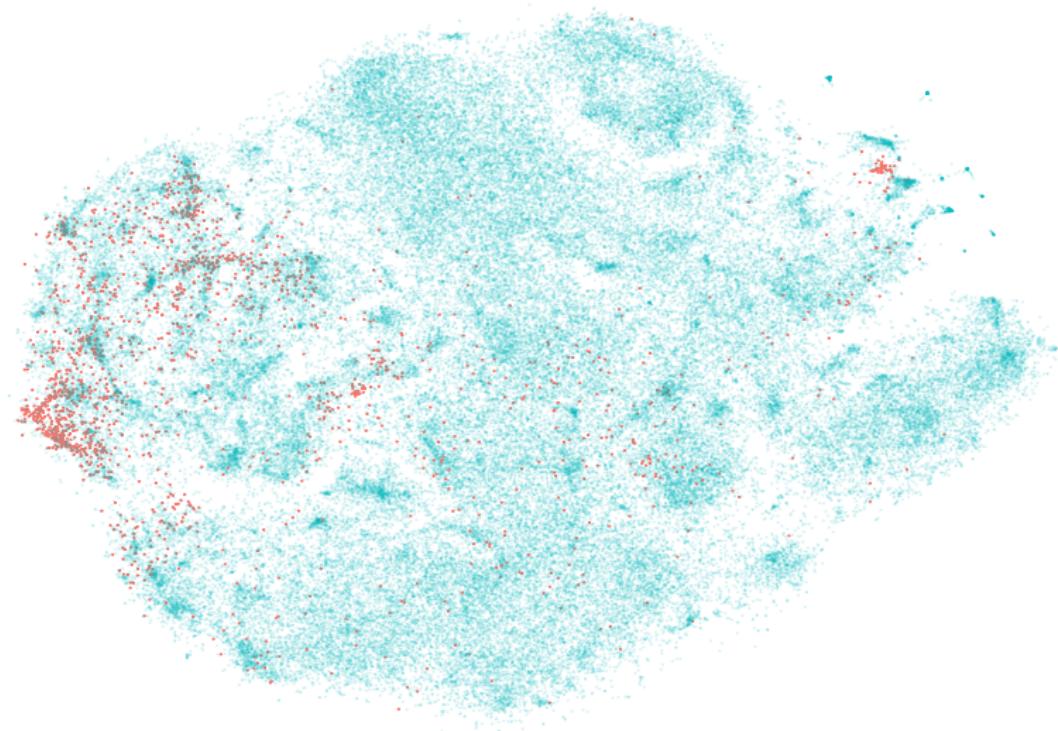
Ландшафт лингвистических исследований



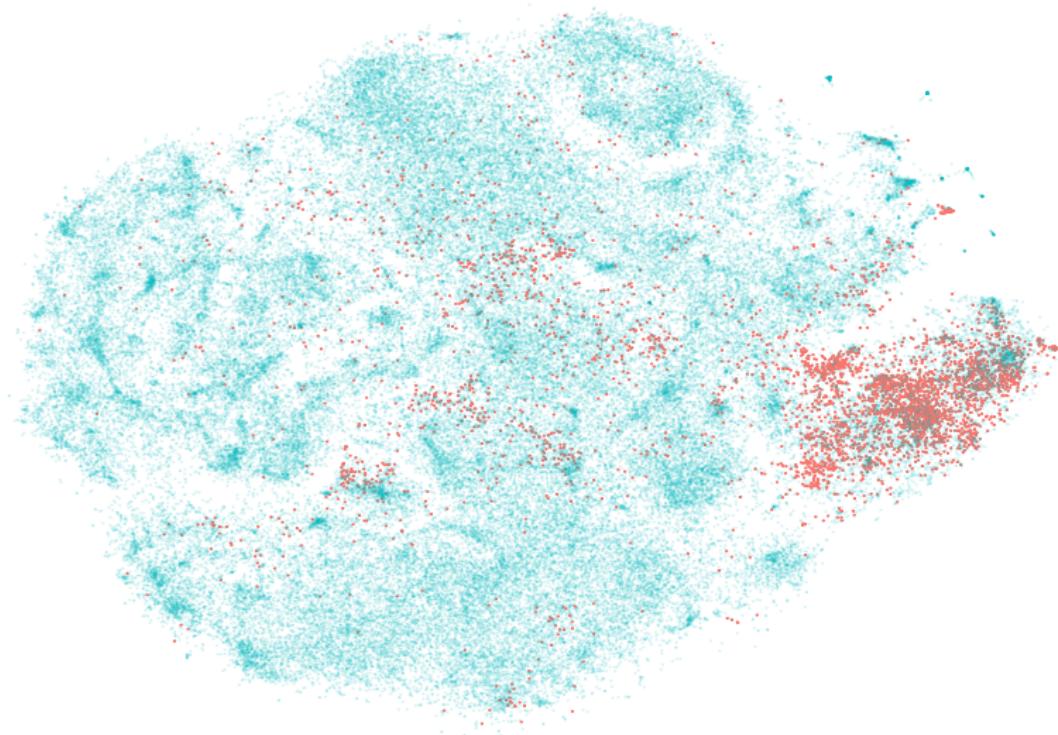
Фонетика и фонология



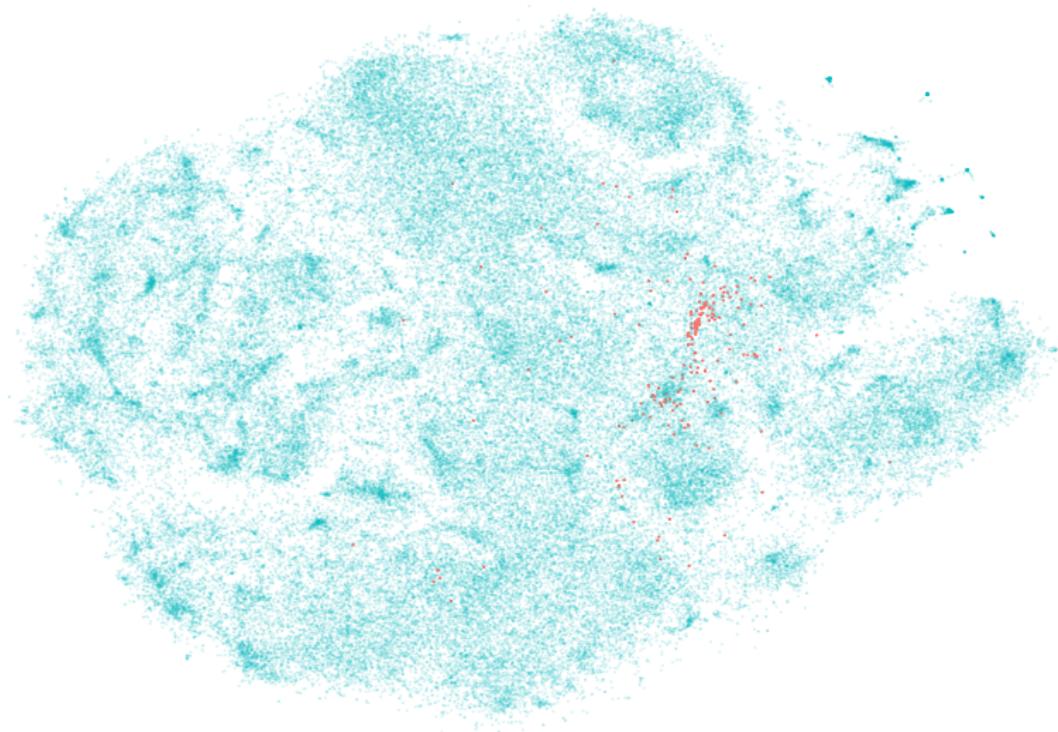
FOLIA PHONIATRICA ET LOGOPAEDICA; JOURNAL OF PHONETICS; JOURNAL OF THE INTERNATIONAL PHONETIC ASSOCIATION; LABORATORY PHONOLOGY; LOGOPEDICS PHONIATRICS VOCOLOGY; PHONETICA; PHONOLOGY; STUDIES IN HISPANIC AND LUSOPHONE LINGUISTICS

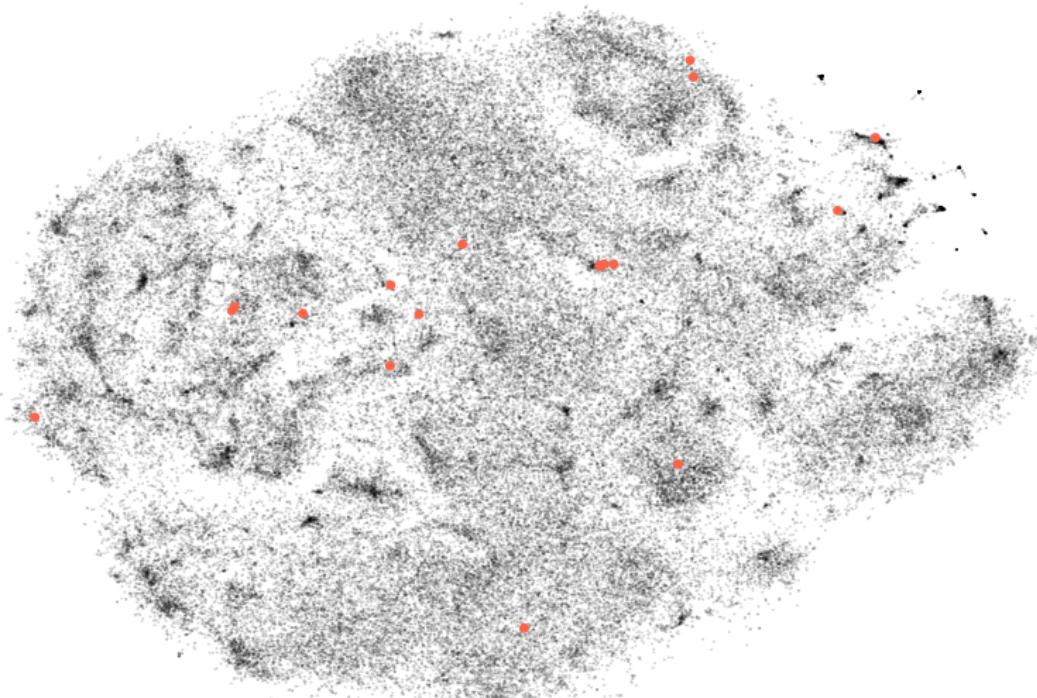


Corpus/comput/quant

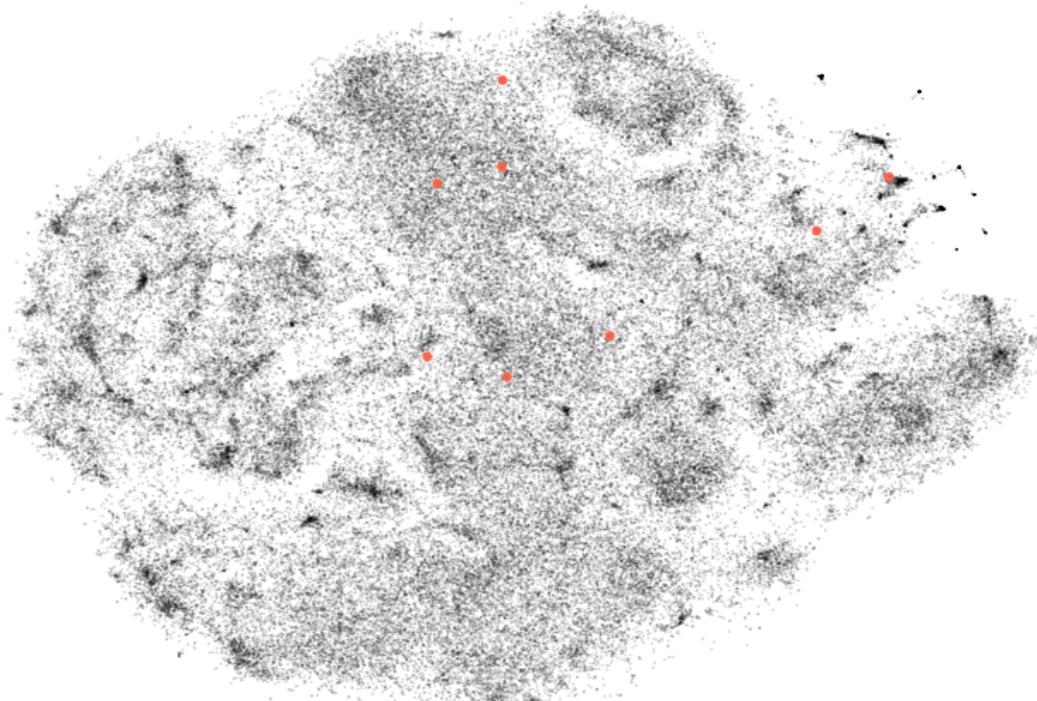


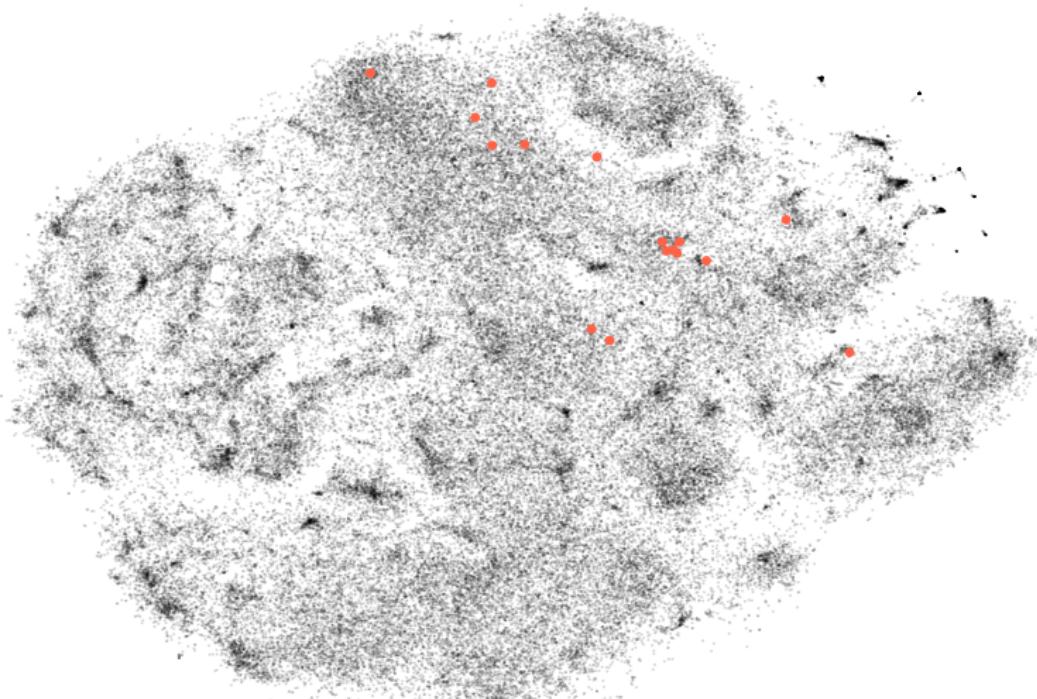
Ономастика





Michael Halliday





План доклада

Введение

Наш проект

Сбор данных

Описательная статистика

Ландшафт лингвистических исследований

Заключение

Заключение

- Нужно исследовать полученное пространство

Заключение

- Нужно исследовать полученное пространство
- Нужно разметить аннотации, а не идти от журналов

Заключение

- Нужно исследовать полученное пространство
- Нужно разметить аннотации, а не идти от журналов
- Разметка журналов нуждается в дополнительной проверке

Заключение

- Нужно исследовать полученное пространство
- Нужно разметить аннотации, а не идти от журналов
- Разметка журналов нуждается в дополнительной проверке
- Интересно посмотреть другие векторизаторы (BERT, RoBERTa)

Заключение

- Нужно исследовать полученное пространство
- Нужно разметить аннотации, а не идти от журналов
- Разметка журналов нуждается в дополнительной проверке
- Интересно посмотреть другие векторизаторы (BERT, RoBERTa)
- Интересно сравнить разные методы уменьшения размерности (t-SNE vs UMAP)

Заключение

- Нужно исследовать полученное пространство
- Нужно разметить аннотации, а не идти от журналов
- Разметка журналов нуждается в дополнительной проверке
- Интересно посмотреть другие векторизаторы (BERT, RoBERTa)
- Интересно сравнить разные методы уменьшения размерности (t-SNE vs UMAP)
- Не стоит обобщать знания, полученные на основе журналов на всю лингвистику: еще бывают книги, главы в сборниках и т. п.

Спасибо за внимание!

Литература

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.

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.

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.

Литература

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.

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

Литература

- 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.
- 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.
- 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.