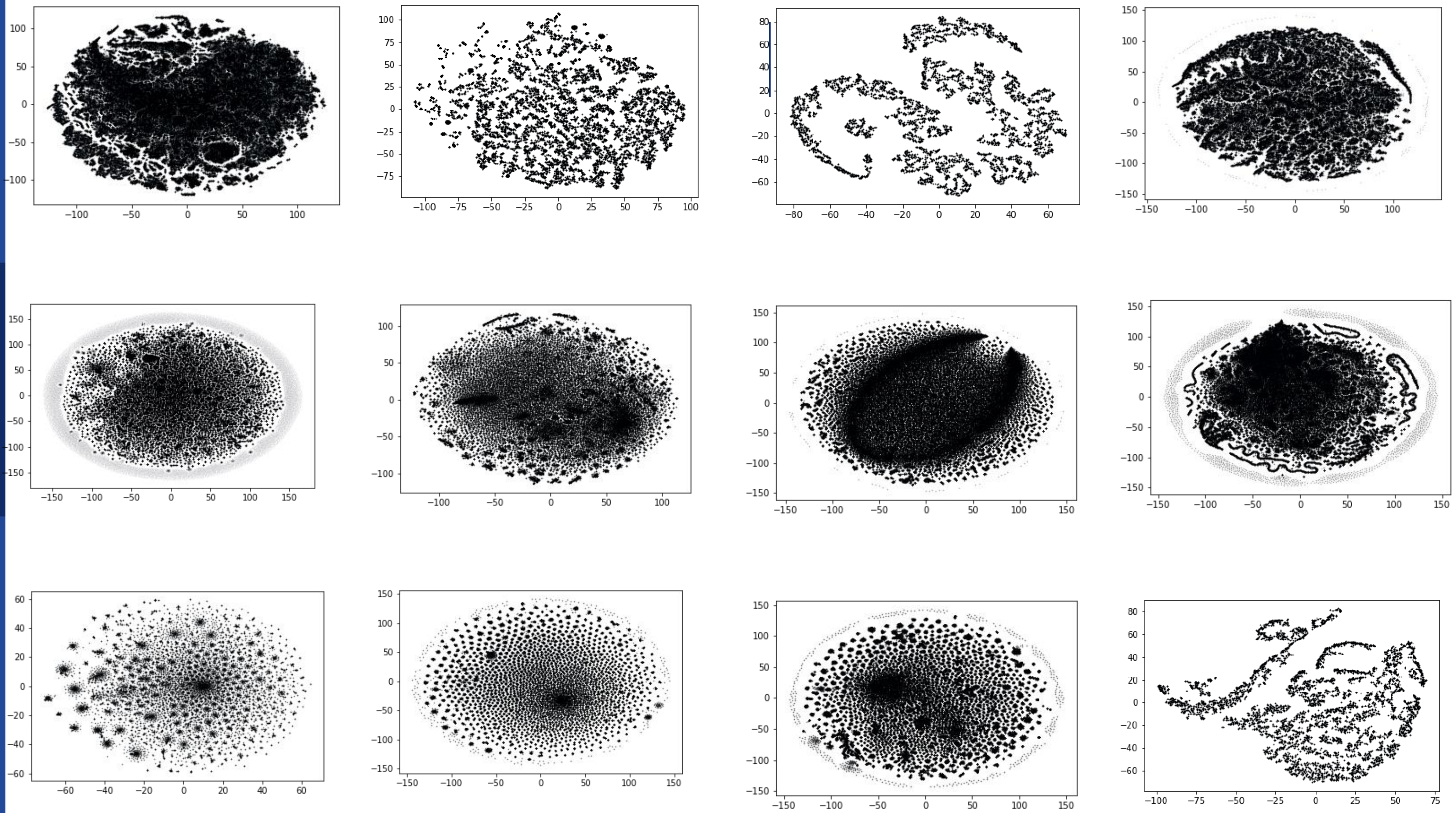
Факультет компьютерных наук

Лаборатория анализа семант

Третья осенняя Сретенская конференция,
Москва, 17-18 ноября 2025

Крупномасштабная структура естественного языка

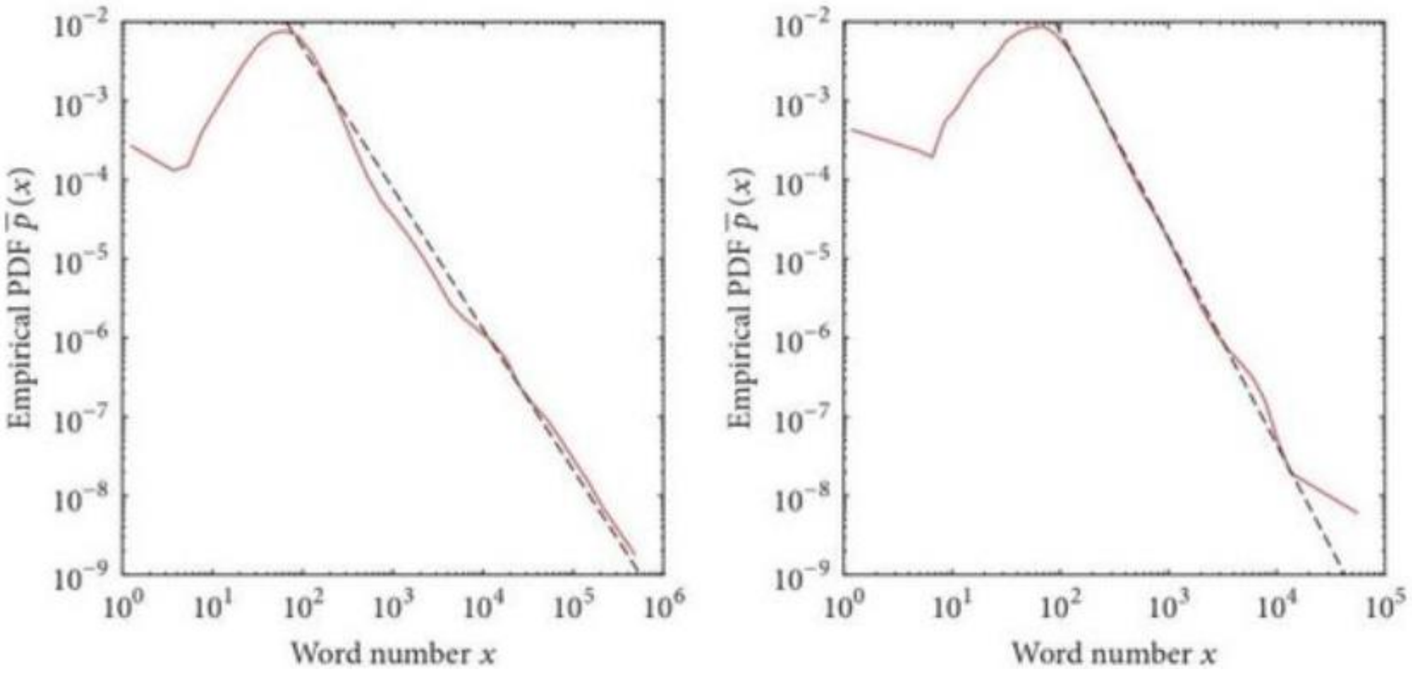
Громов Василий Александрович
проф. департамента анализа данных и искусственного интеллекта НИУ ВШЭ,
руководитель Лаборатории анализа семантики Центра языковых и семантических технологий НИУ ВШЭ
+79036865072, stroller@rambler.ru



100 50 0 −50 −100
−100 −50 0 50 100
100 75 50 25 0 −25 −50 −75
−100 −75 −50 −25 0 25 50 75 100
80 60 40 20 0 −20 −40 −60
−80 −60 −40 −20 0 20 40 60
150 100 50 0 −50 −100 −150
−150 −100 −50 0 50 100
150 100 50 0 −50 −100 −150
−150 −100 −50 0 50 100 150
100 50 0 −50 −100
−100 −50 0 50 100
150 100 50 0 −50 −100 −150
−150 −100 −50 0 50 100 150
150 100 50 0 −50 −100 −150
−150 −100 −50 0 50 100 150
60 40 20 0 −20 −40 −60
−60 −40 −20 0 20 40 60
150 100 50 0 −50 −100 −150
−150 −100 −50 0 50 100 150
150 100 50 0 −50 −100 −150
−150 −100 −50 0 50 100 150
80 60 40 20 0 −20 −40 −60
−100 −75 −50 −25 0 25 50 75

Язык — это самоорганизующаяся критическая система

- Язык — это **самоорганизующаяся критическая система**.
- Тексты языка представляют собой «лавины».
- Количество слов в текстах подчиняется степенному закону распределения.
- $p(x) = ax^{-\tau},\ x > x_{min}$



Эмпирическая функция плотности вероятности для сл. величины "количество слов в тексте" (лог-лог шкала). Русские литературные произведения (слева) и поэзия (справа)
V. A. Gromov, A. M. Migrina. A Language as a Self-organized Critical System. Complexity 2017.

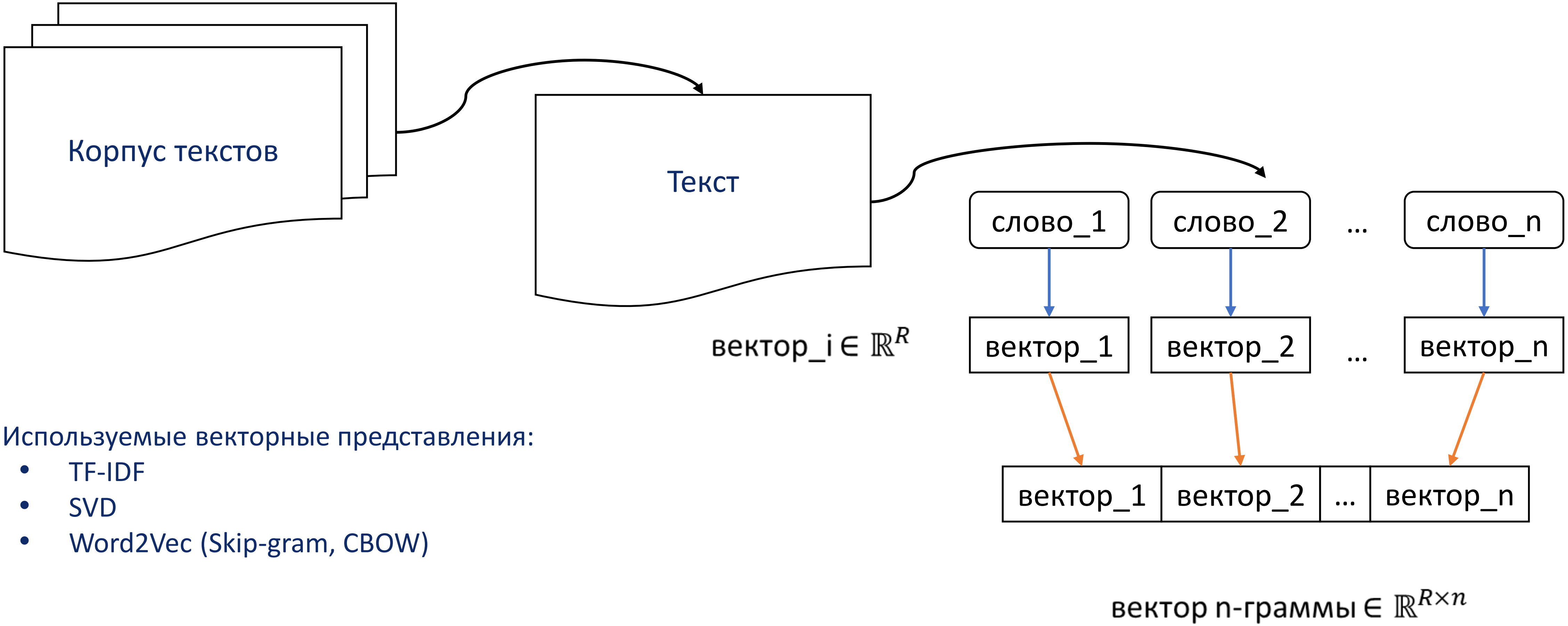
Язык — это самоорганизующаяся критическая система

	Английский				Русский			
	Оценки по коллапсу данных		Оценки по критерию К-С		Оценки по коллапсу данынх		Оценки по критерию К-С	
Тип текста	(nw)	(dl)	(nw)	(dl)	(nw)	(dl)	(nw)	(dl)
Литература	114	1.30	55	1.29	72	1.78	56	1.97
Поэзия	125	1.91	109	1.89	112	2.61	98	2.63
Социальные сети	111	2.65	124	2.63	183	2.10	151	2.15
Объединённые тексты	109	1.39	42	1.34	85	1.82	62	1.95

Таблица 1: Оценка показателей степенной зависимости и нижние границы распределения количества слов в текстах литературных произведений и социальных сетях, nw – число слов, dl – безразмерные

	Английский				Русский			
	Оценки по коллапсу данных		Оценки по критерию К-С		Оценки по коллапсу данных		Оценки по критерию К-С	
Период	(nw)	(dl)	(nw)	(dl)	(nw)	(dl)	(nw)	(dl)
19ый век и раннее	118	1.37	75	1.37	91	1.71	53	1.87
20ый век	121	1.19	54	1.21	74	1.75	53	2.03

Таблица 1: Оценка показателей степенной зависимости и нижние границы распределения количества слов в текстах литературных произведений 19-го (и раннее) и 20-го века; nw – число слов, dl – безразмерные



Используемые векторные представления:

- TF-IDF
- SVD
- Word2Vec (Skip-gram, CBOW)

Модели, методы и алгоритмы

Модель оценки размерности Хаусдорфа

Для построения данной модели используется теорема о связи суммы степеней длин рёбер остовного дерева минимального веса с внутренней размерностью топологического пространства:

$$E_0^\alpha(x_1, \dots, x_n) = \mathrm{C}(d, \alpha) \cdot n^{\frac{d-\alpha}{d}},$$

где:

$$E_0^\alpha(x_1, \dots, x_n) = \sum_{e \in T_n(\{x_n\}_{n\in\mathbb{N}})} |e|^\alpha,$$

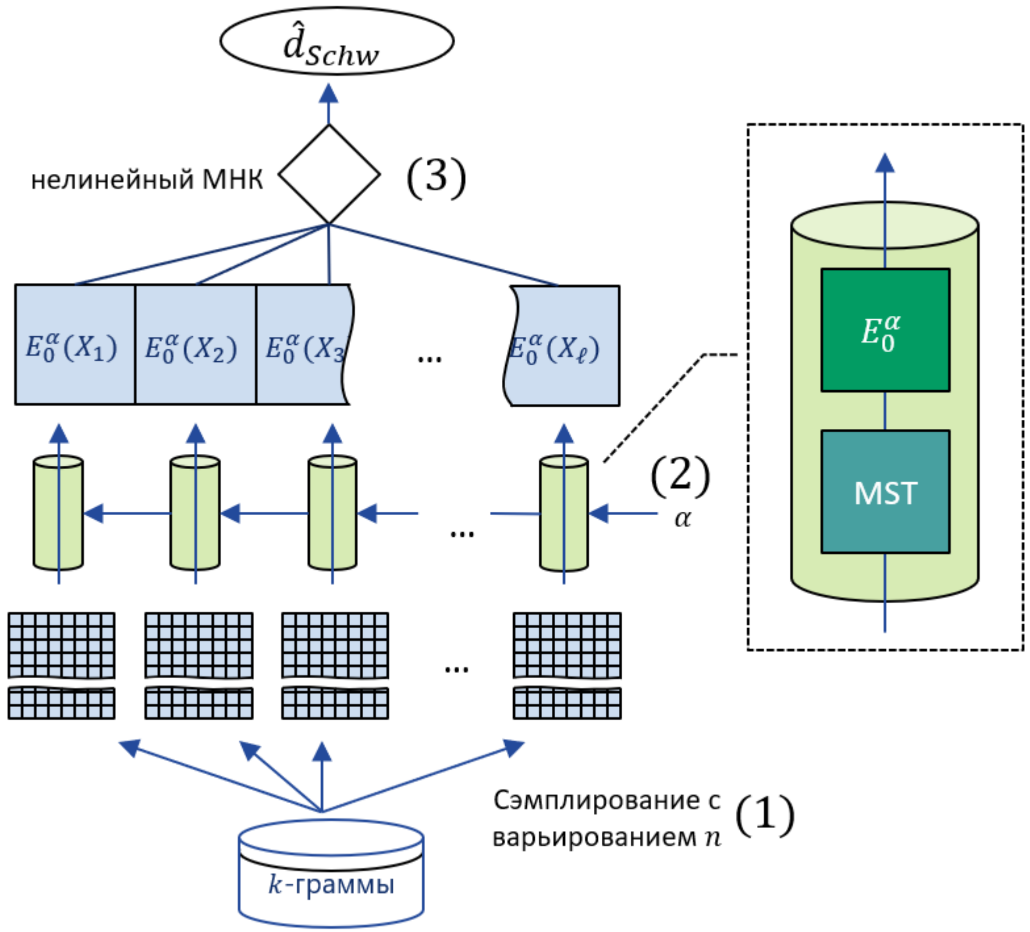
Здесь и далее $T_n(\{x_n\}_{n\in\mathbb{N}})$ – остовное дерево минимального веса (MST) над множеством точек (k-грамм) $\{x_n\}_{n\in\mathbb{N}}$, n – общее число всех k-грамм.

Линеаризация уравнения из теоремы позволяет решить задачу методом наименьших квадратов (МНК):

$$\ln\left(E_0^\alpha(x_1, \dots, x_n)\right) = \ln\left(\mathrm{C}(d, \alpha)\right) + \frac{d-\alpha}{d}\ln(n)$$

$$f(\ln(n); \theta) = \theta_0 + \theta_1 \cdot \ln(n)$$

$$\hat{d}_{Schw} = \frac{\alpha}{1 - \hat{\theta}_1}$$



Модели, методы и алгоритмы

Модель оценки топологической размерности

Идея заключается в использовании статистики $M_n(d)$:

$$M_n(d) := \frac{1}{n} \sum_{X_i \in T(\{x_n\}_{n\in\mathbb{N}} \subset \mathcal{X}_d)} \left(\deg(X_i)\right)^2.$$

Известно, что M_n монотонно не убывает с ростом d.

Для каждой гипотетически возможной истинной размерности i рассчитывается $f_{prior}(M_n|d^* = i) \approx N(\hat{\mu}_i, \hat{\sigma}_i^2)$, где:

$$\hat{\mu}_i := \frac{1}{L} \sum_{j=0}^{L} M\left(i; \{x_n\}_j \sim U(\mathbf{0}_i, \mathbf{1}_i)\right),$$

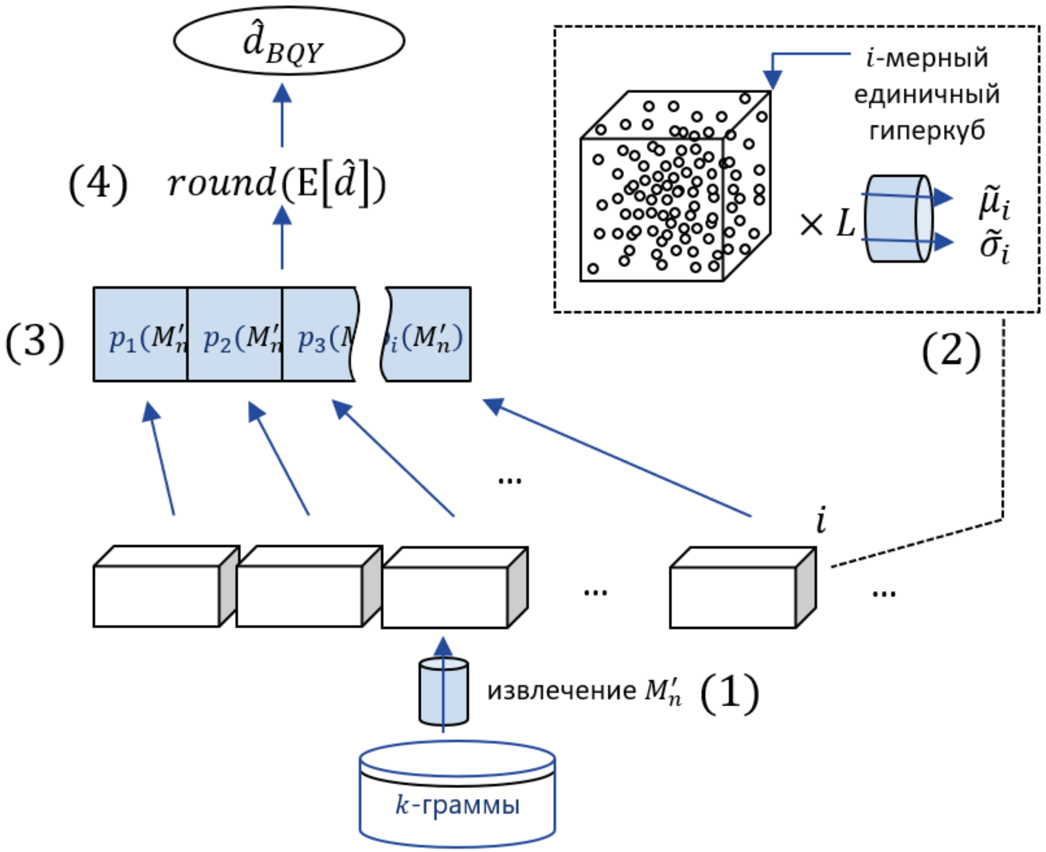
$$\hat{\sigma}_i^2 := \frac{n}{L-1} \sum_{j=0}^{L} \left(M\left(i; \{x_n\}_j \sim U(\mathbf{0}_i, \mathbf{1}_i)\right) - \hat{\mu}_i\right)^2.$$

Здесь L – число синтетических выборок, по n точек в каждой, выбранных из равномерного распределения в i-мерном гиперкубе.

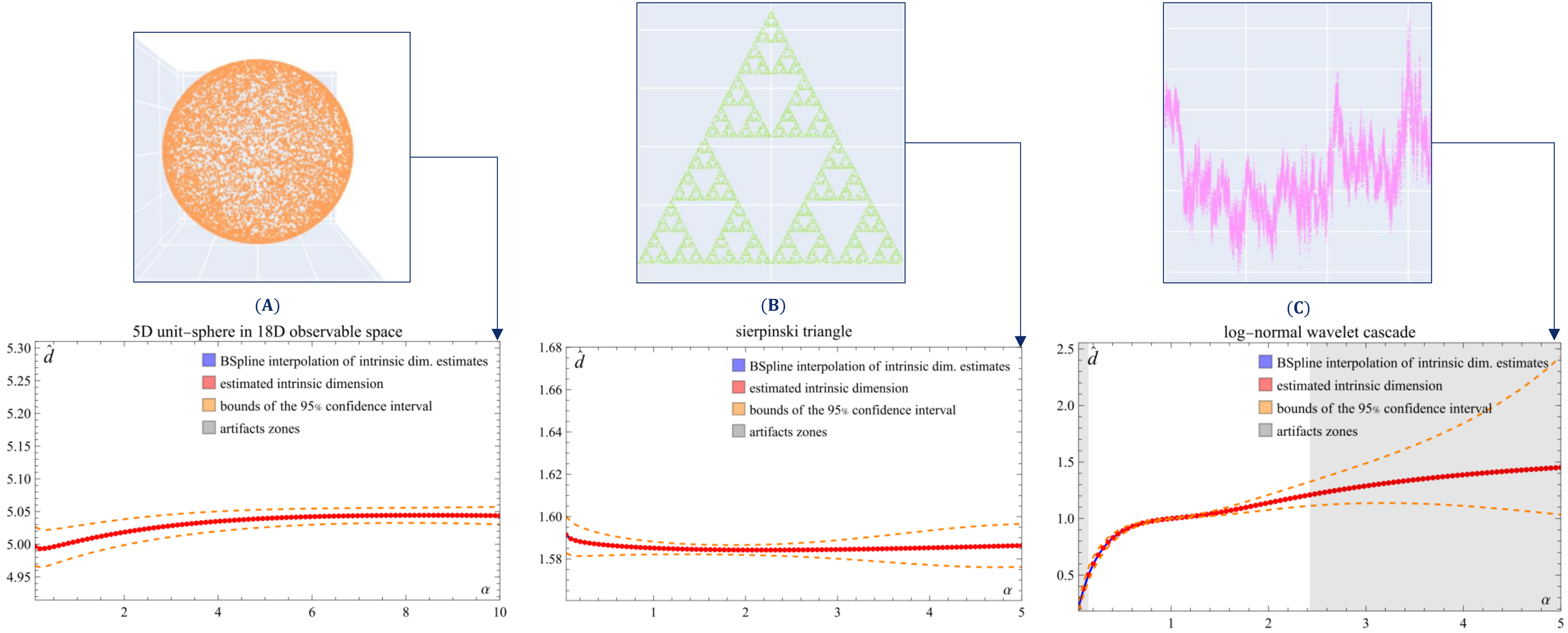
Тогда по теореме Байеса можем рассчитать:

$$p\{\hat{d} = i\} = p(i|M_n') = \frac{N\left(M_n'; \hat{\mu}_i, \frac{\hat{\sigma}_i^2}{n'}\right)}{\sum_j N\left(M_n'; \hat{\mu}_j, \frac{\hat{\sigma}_j^2}{n'}\right)};$$

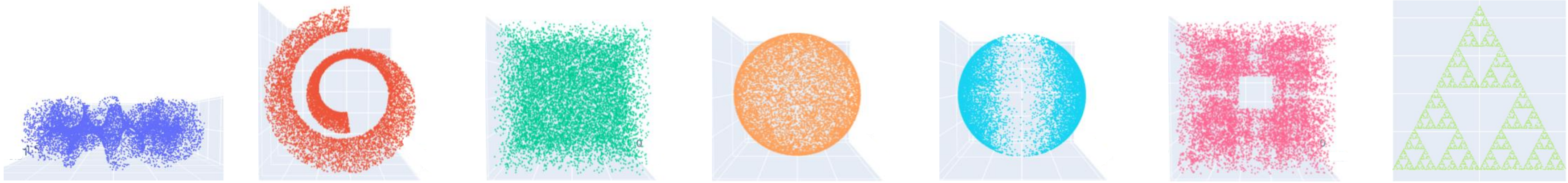
$$\hat{d}_{BQY}(M_n') = round(E[\hat{d}]) = round\left(\sum_i p(i|M_n') * i\right)$$



Вычислительный эксперимент: тест моделей оценки размерности

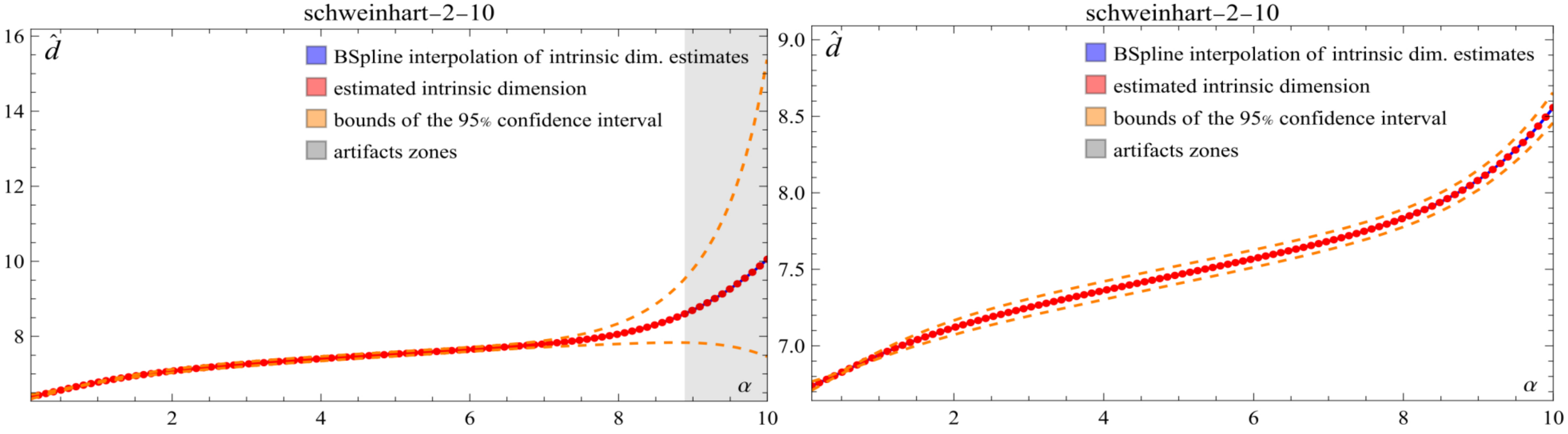


Вычислительный эксперимент: тест моделей оценки размерности



α	Mobius tape APE	Swiss-roll APE	Unit-cube APE	Unit-sphere APE	Unit-sphere (Gauss.) APE	Menger sponge APE	Sierpinski carpet APE
1.00	0.47%	0.43%	1.35%	0.07%	0.68%	1.56%	0.07%
2.00	0.77%	0.43%	1.57%	0.03%	1.26%	1.68%	0.03%
3.00	1.15%	0.47%	1.77%	0.03%	2.41%	1.76%	0.11%
4.00	1.59%	0.55%	1.98%	0.08%	3.79%	1.82%	0.18%
5.00	2.10%	0.68%	2.20%	0.12%	5.02%	1.85%	0.24%
6.00	2.66%	0.85%	2.43%	0.16%	5.99%	1.87%	0.28%
7.00	3.22%	1.08%	2.69%	0.19%	6.77%	1.90%	0.32%
8.00	3.72%	1.36%	2.97%	0.22%	7.43%	1.92%	0.35%
9.00	4.11%	1.69%	3.28%	0.25%	8.05%	1.97%	0.38%
10.00	4.33%	2.04%	3.62%	0.29%	8.58%	2.03%	0.42%
mean	**2.41%**	**0.96%**	**2.38%**	**0.14%**	**5.00%**	**1.84%**	**0.24%**

Внутренние размерности естественных языков



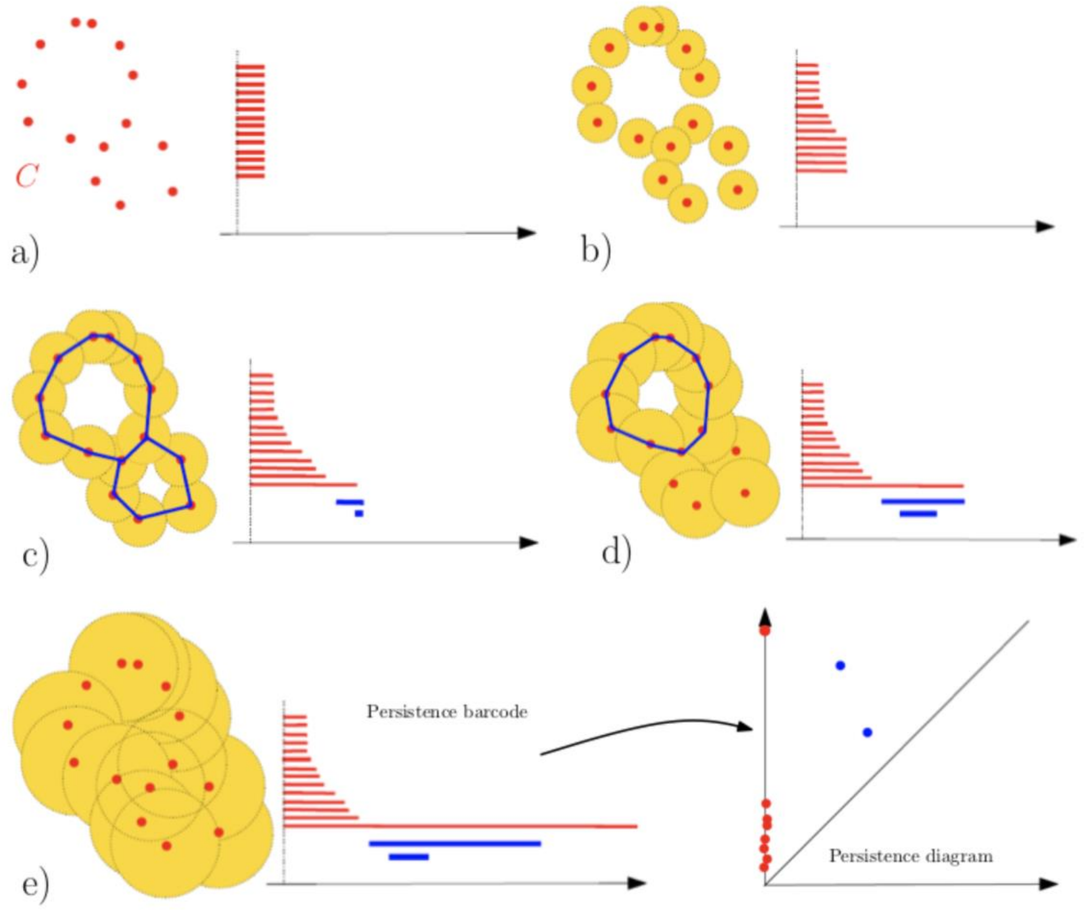
Зависимость оценок алгоритма Швайнхарта от параметра α для CBOW-представлений национальной литературы ($n = 2, d = 10$): (A) русского языка; (B) английского языка. Красные точки являются оценками алгоритма, синяя линия представляет собой B-сплайн интерполяцию зависимости оценок от α, оранжевый пунктир – верхнюю и нижнюю границы 95% доверительного интервала. На рисунке также представлены критерии допустимости параметра α, такие как расхождение доверительных интервалов (серые зоны).

Топологический анализ данных

Выделение топологической структуры семантического пространства

Поиск:

1. Области сгущения, содержащих много n-грамм
2. «Дырки» – области, n-граммы откуда не употребляются или употребляются крайне редко.



Выделение гомологии на участке векторного пространства

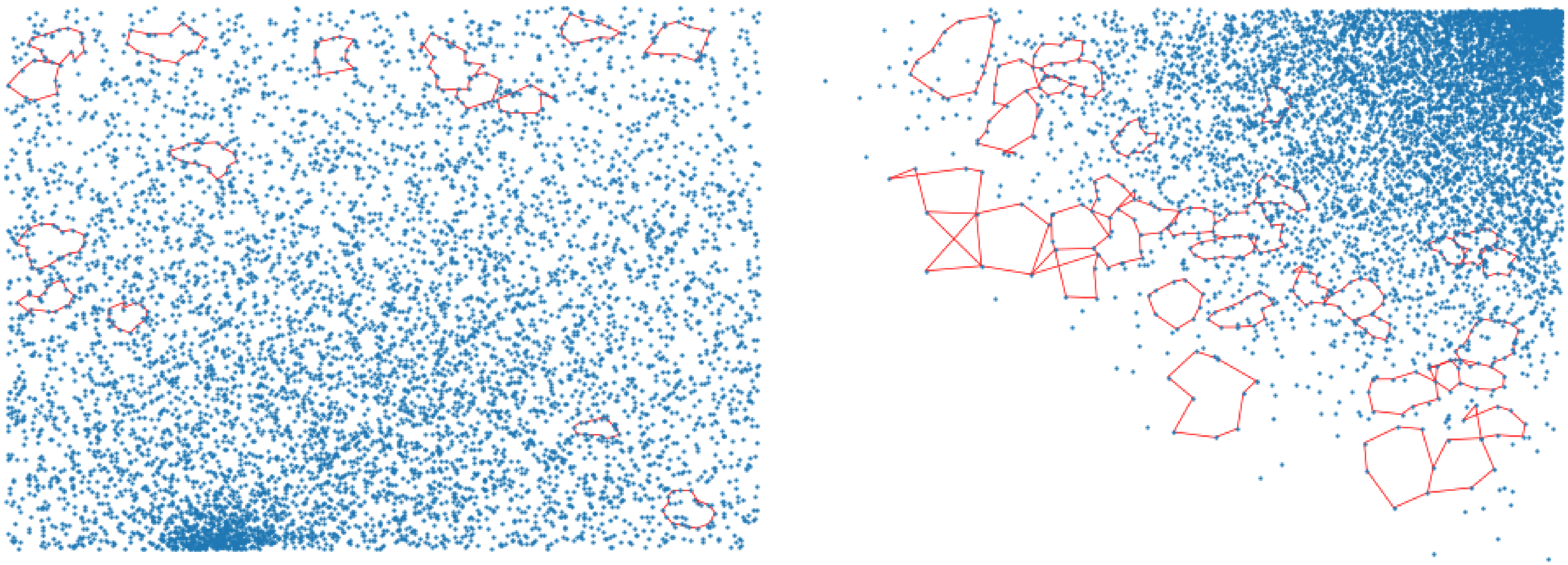


Рис.: Гомологии в центре (слева) и на крае (справа) векторного пространства слов русского языка (+PCA).

Выделение гомологий в векторном пространстве слов, биграмм и триграмм

Фильтрация гомологий по:

- Персистентности (0.99 квантиль)
- Диаметру (больше диаметра групп синонимов)

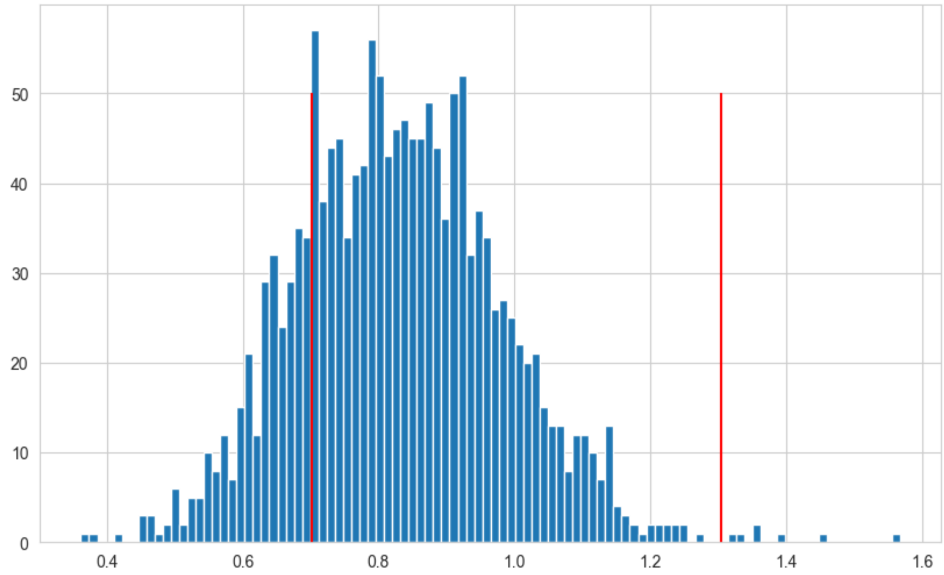


Рис.: Распределение диаметров гомологий первого порядка в векторном пространстве слов русского языка. Красными линиями отмечены среднее и максимальное значения диаметров групп синонимов.

Выделение гомологий в векторном пространстве слов, биграмм и триграмм

	Слова	Биграммы	Триграммы
Русский	7	12	12
Английский	4	5	7

Таблица: Число отобранных гомологий первого порядка

смешливый, черноглазый, чернобровый, молодица, баять, вишь, сякой, срамить, бранить, сердиться, конфузиться, сконфузить, смущенный, растерянный, озадаченный, ошарашивать, оторопеть, озадаченно, многозначительно, благожелательно, доброжелательно, дружелюбный, приветливый, добродушный, простоватый, разбитной, бойкий

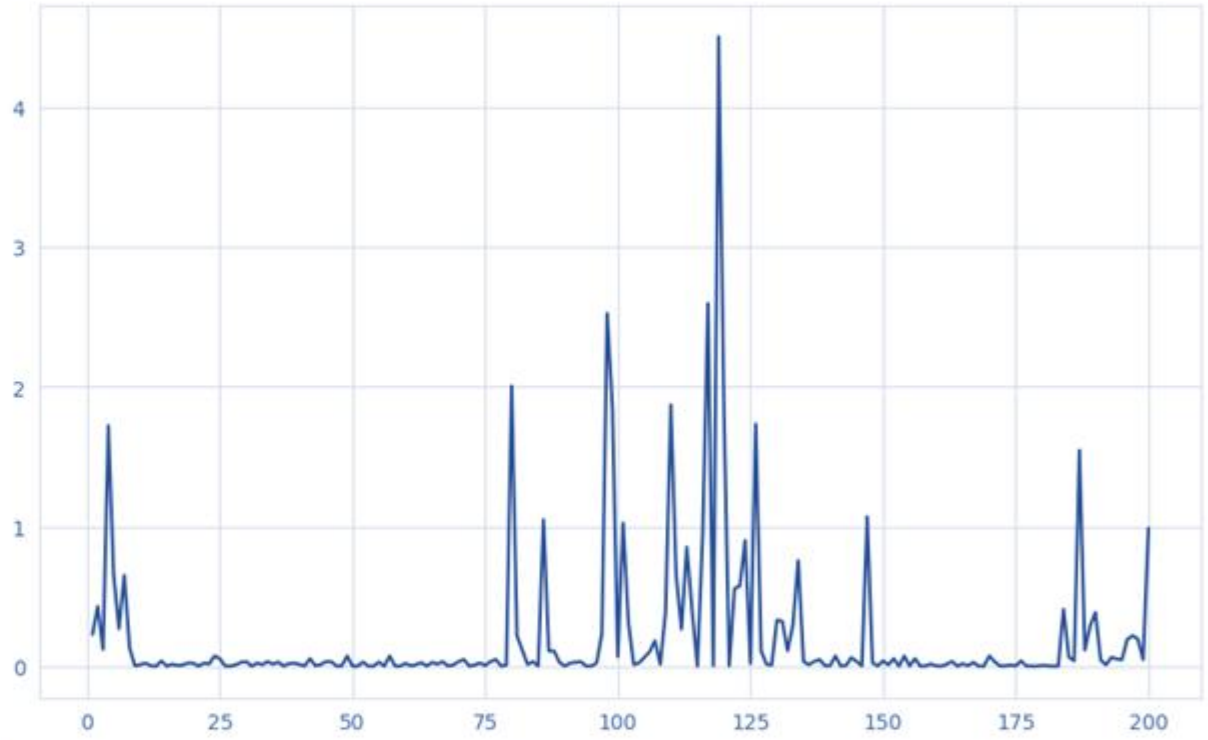
ладиться княжество, петровский империя, китайский государство, немецкий власть, латвийский власть, говорить могущество, столько могущество, столько теплота, сколько любовь, ревнивость любовь, выглядеть любовь, выглядеть жизнь, выглядеть либо, озорница либо, эмоциональный либо, эмоциональный будущее, руководимый будущее, многий будущее, многий предок, всеведущий потомок

Временные ряды

Текст как многомерный хаотический **временной ряд**

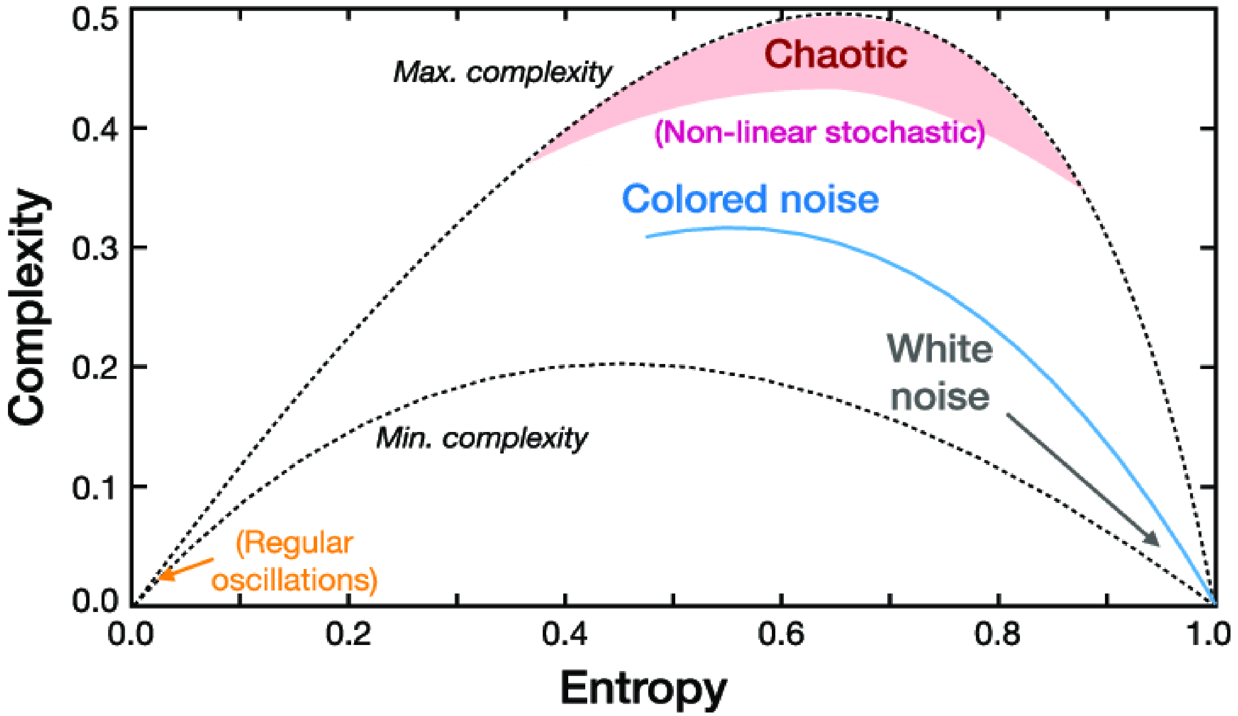
Характеристики:

- Энтропия-сложность
- Размерности странных аттракторов



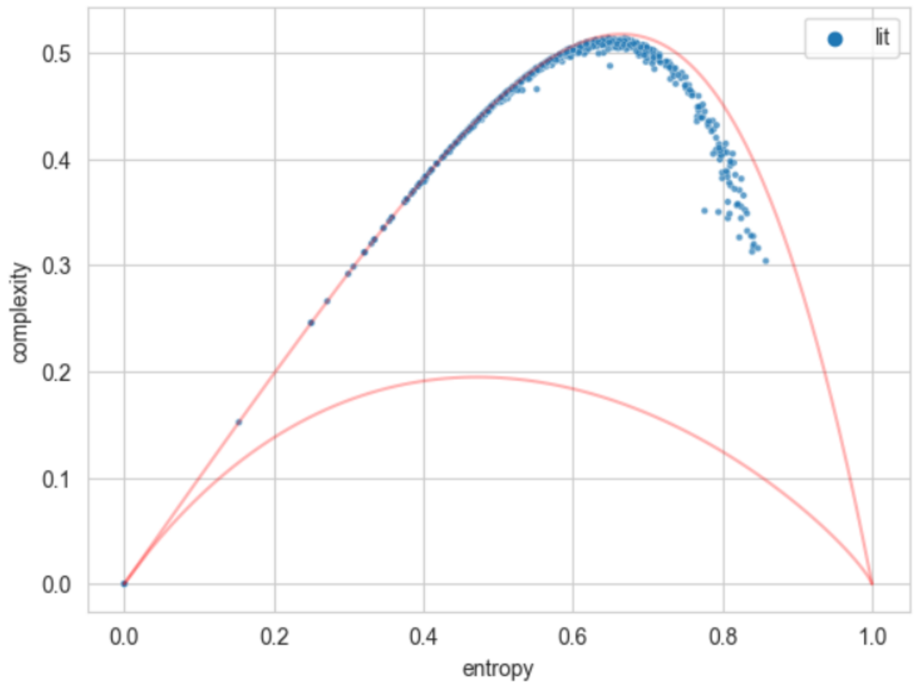
Визуализация первой компоненты временного ряда первых 200 слов романа «Война и мир» Л. Н. Толстого

Энтропия-сложность

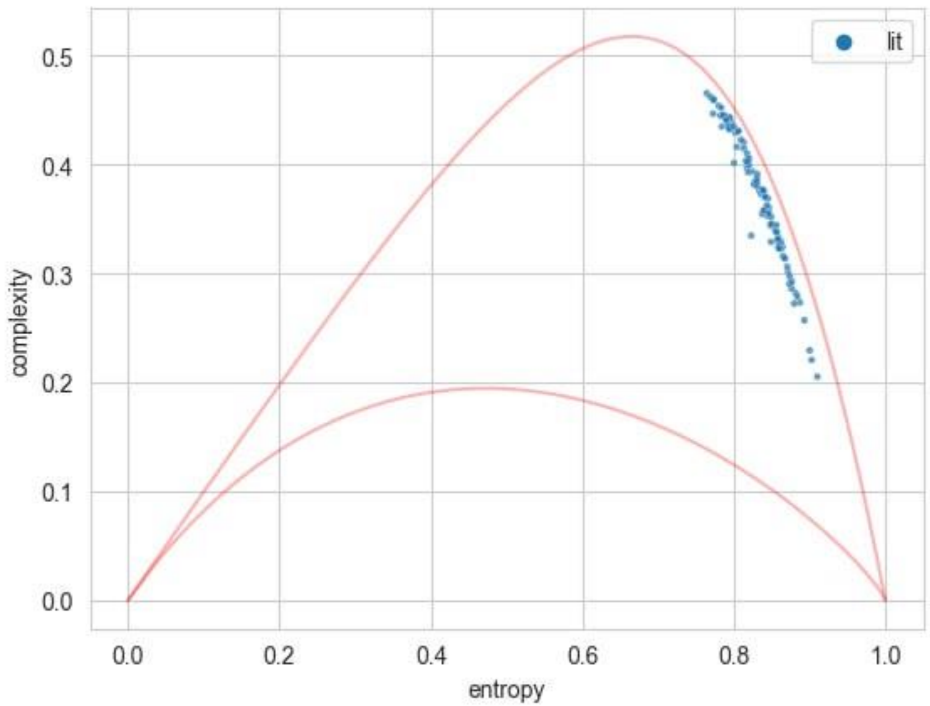


From paper: O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino and M. A. Fuentes, “Distinguishing noise from chaos,” Physical review letters, vol. 99, no. 15, 154102, 2007

Энтропия-сложность

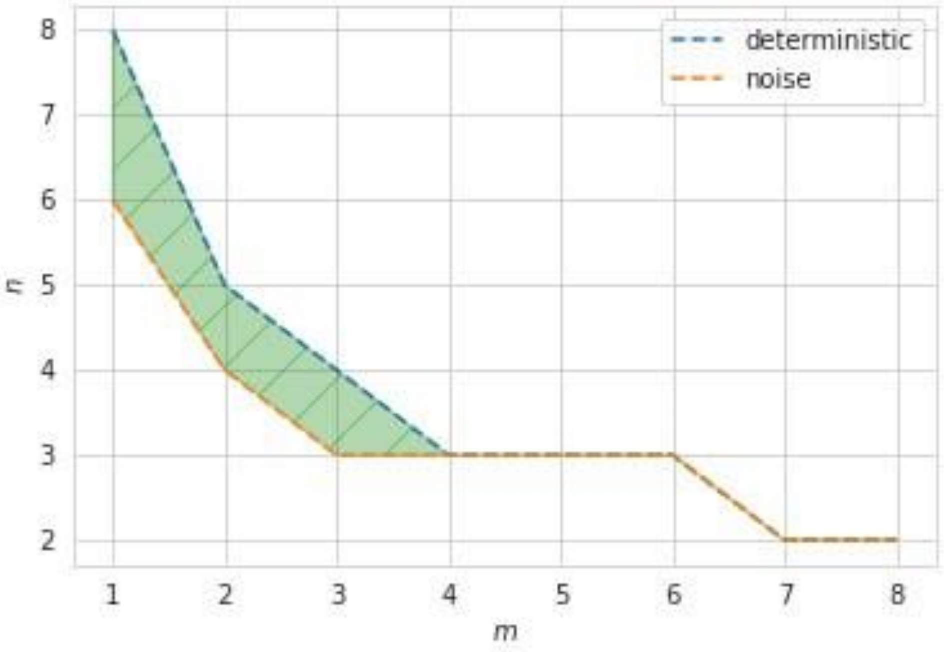


Русский язык

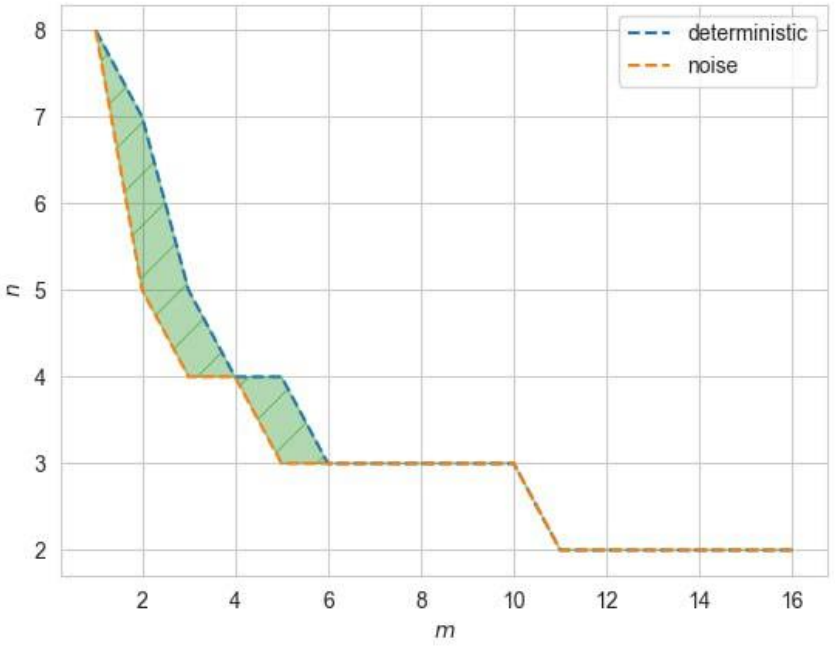


Белорусский язык

Области хаоса

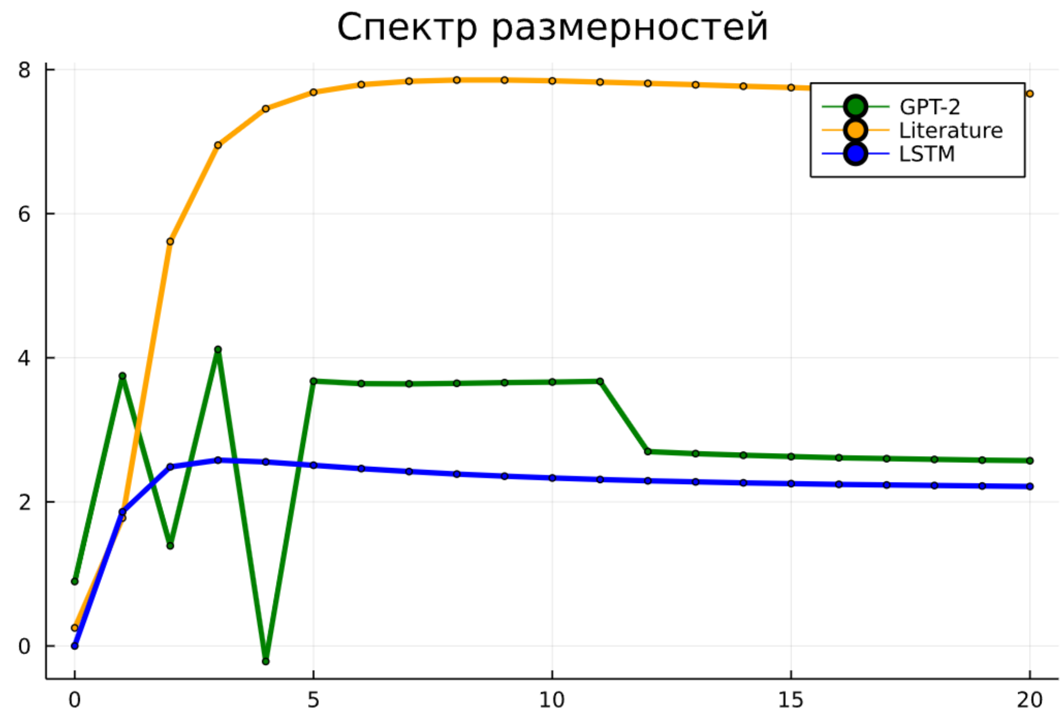


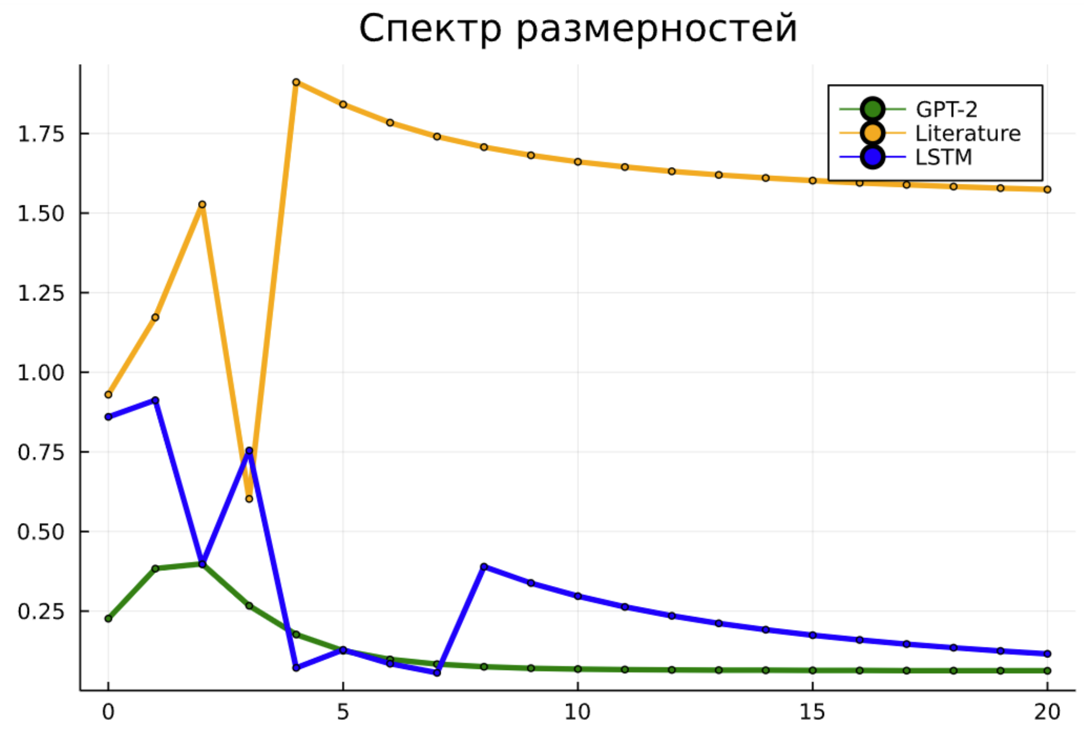
Русский язык



Белорусский язык

Обобщённая размерность странного аттрактора





Идентификация ботов

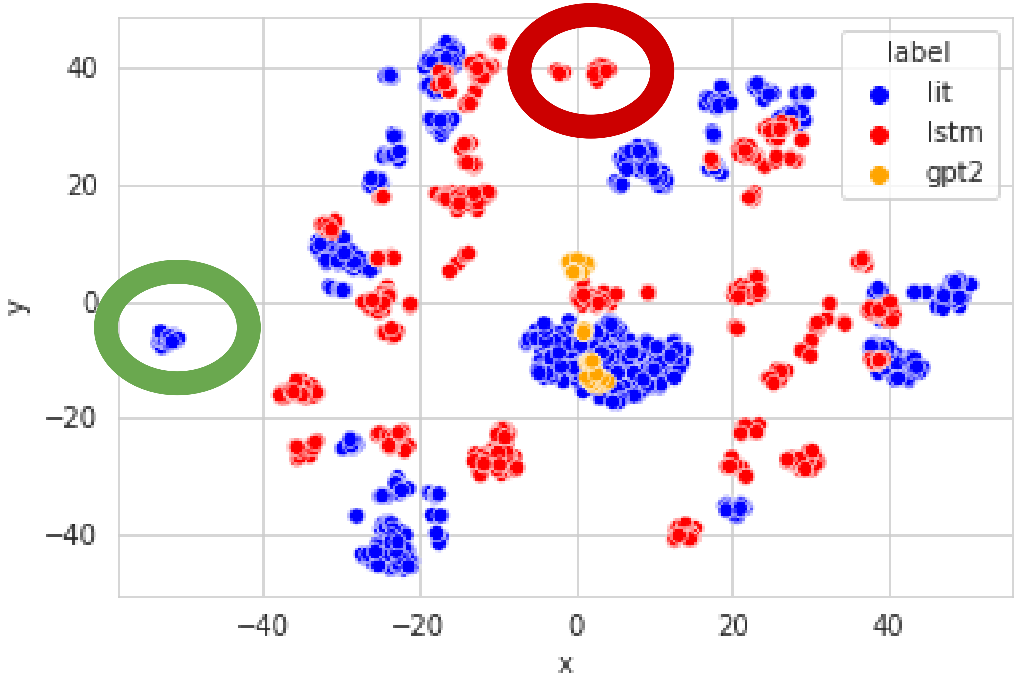
Боты – технология искусственного интеллекта, позволяющая генерировать **тексты**.

Тексты ботов могут:

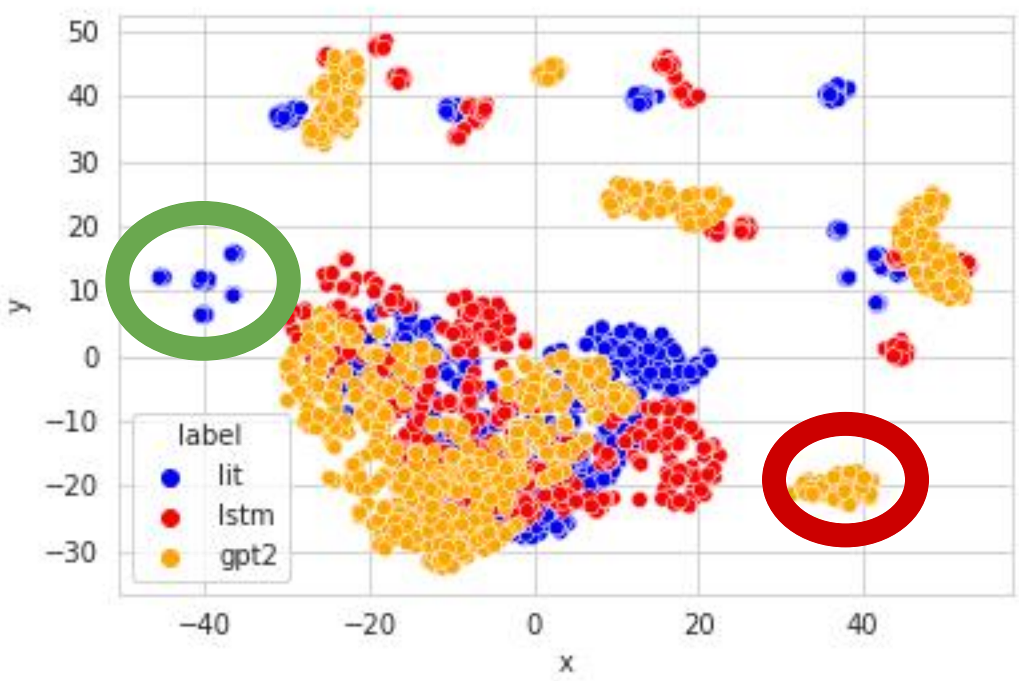
- Вводить в заблуждение **своими текстами**
- Оказывать негативное влияние на **развитие языковой личности**

На данный момент **не существует** систем, отличающих большее число ботов.

TSNE-визуализация семантического пространства (Хаилонакея)



Язык: Русский
Векторное представление: CBOW



Язык: Русский
Векторное представление: SVD

Выделение гомологий в векторном пространстве слов, биграмм и триграмм

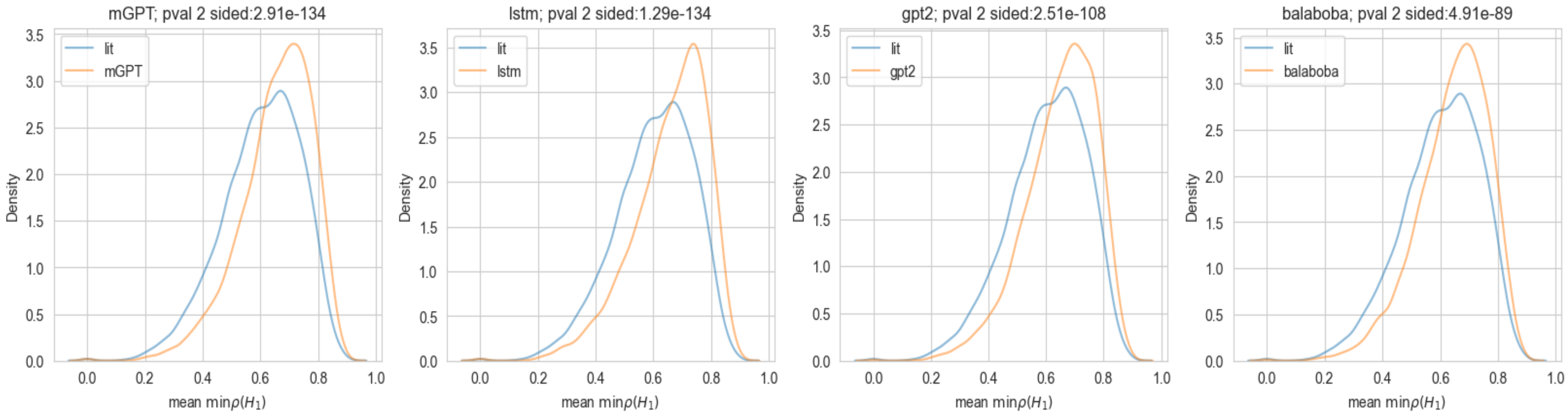


Рис.: Распределение расстояний до ближайших гомологий первого порядка в векторном пространстве слов русского языка. Синим показано распределение для текстов литературы, оранжевым – ботов (слева направо: mGPT, LSTM, GPT-2, YaLM).

Классификация по внутрикластерным расстояниям

Языки	Русский			Английский			Вьетнамский		
Тексты, написанные людьми, vs	Все боты	LSTM	GPT-2	Все боты	LSTM	GPT-2	Все боты	LSTM	GPT-3
K-Means	0.934/0.934 0.934/0.934	1.0	0.916/0.916 0.916/0.916	0.974/0.969 0.976/0.975	1.0	0.881/0.0.834 0.895/0.881	0.903/0.883 0.883/0.883	1.0	0.881/0.911 0.911/0.911
Wishart	0.954/0.954 0.954/0.954	1.0	0.944/0.944 0.945/0.945	0.975/0.968 0.976/0.975	1.0	0.881/0.834 0.895/0.881	0.896/0.897 0.900/0.896	1.0	0.900/0.896 0.897/0.896
C-Means	0.894/0.893 0.892/0.894	1.0	0.857/0.856 0.865/0.857	0.970/0.959 0.971/0.970	1.0	0.921/0.883 0.848/0.921	0.893/0.890 0.892/0.890	1.0	0.871/0.872 0.874/0.871
Wishart на нечётких числах	0.913/0.913 0.913/0.913	1.0	0.911/0.911 0.911/0.911	0.947/0.933 0.952/0.947	1.0	0.940/0.926 0.944/0.940	0.942/0.942 0.943/0.942	1.0	0.926/0.926 0.926/0.926

Метрики классификации: accuracy/F-score (weighted average); Precision/Recall (weighted average)

Введение нечёткости повышает точность различения текстов английского и вьетнамского языков

Семинар “Нелинейные явления в человекомерных системах”
Москва, Покровский бульвар, 11, S902

+79036865072, stroller@rambler.ru

СПАСИБО ЗА ВНИМАНИЕ

ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ