Студенческий научный электронный журнал
StudArctic Forum

ПЕТРОЗАВОДСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Т. 8, № 2. С. 30—35 **Компьютерные науки и информатика** **2023**
Научная статья
УДК 81`32

КУУСЕЛА Демид Александрович
бакалавриат, Петрозаводский государственный университет (Петрозаводск, Российская Федерация)
demid.kuusela@gmail.com

АНАЛИЗ ЛЕКСИЧЕСКИХ СПЕКТРОВ ТЕКСТОВ С ПОМОЩЬЮ МАТЕМАТИЧЕСКИХ МЕТОДОВ

Научный руководитель:
Москин Николай Дмитриевич
Статья поступила: 09.06.2023;
Принята к публикации: 17.06.2023;
Размещена в сети: 25.06.2023

Аннотация. В статье рассматривается анализ лексических спектров текстов с помощью математических методов и компьютерных технологий. В частности, описывается аппроксимация значений лексических спектров текстов с помощью логарифмической кривой.
Ключевые слова: атрибуция текстов, установление авторства, лексический спектр, аппроксимация, ИС «СМАЛТ»

Для цитирования: Куусела Д. А. Анализ лексических спектров текстов с помощью математических методов // StudArctic Forum. 2023. Т. 8, № 2. С. 30—35.

Атрибуция — это исследование текстовых произведений с целью определения авторства. Проблема определения авторства связана с существованием анонимных и псевдонимных текстов и представляет собой одну из древнейших филологических задач. В последние десятилетия для решения задач атрибуции все чаще стали применяться математические методы и компьютерные технологии. Это обусловлено тем, что традиционные методы атрибуции, основанные на субъективной оценке экспертов, с большей вероятностью бывают неполными и предвзятыми. Компьютерные технологии также позволяют проводить анализ больших объемов текстов и выявлять статистические закономерности в стилистическом почерке и авторской манере писателей. Например, математические методы и компьютерные технологии использовались для атрибуции текстов Ф. М. Достоевского [Кулаков].

Один из таких методов был разработан в 1976 году Гейром Хьетсо, известным норвежским специалистом по использованию компьютерных технологий для атрибуции литературных произведений. Методика основана на 15 лингвостатистических параметрах [Хьетсо: 31]:

1. Общее распределение частей речи в первых двух и в последних трех позициях предложения.
2. Распределение частей речи в первой позиции предложения.
3. Распределение частей речи во второй позиции предложения.
4. Сочетание частей речи в первых двух позициях предложения.
5. Распределение частей речи в третьей с конца позиции предложения.
6. Распределение частей речи в предпоследней позиции предложения.
7. Распределение частей речи в последней позиции предложения.
8. Сочетание частей речи в последних трех позициях предложения.
9. Средняя длина слова в буквах, вычисляемая на основании выборок размером в 500 текстовых слов.
10. Общее распределение длины слова.
11. Средняя длина предложения в словах, вычисляемая на основании выборок размером в 30 предложений.
12. Общее распределение длины предложения.
13. Лексический спектр текста на уровне словаря.
14. Лексический спектр текста на уровне текста.
15. Индекс разнообразия лексики.

В исследованиях Г. Хьетсо отмечается, что «можно исследовать стилистический почерк писателя и с точки зрения лексического спектра текста, т. е. с точки зрения распределения частот слов в тексте» [Хьетсо: 58]. Для этого Г. Хьетсо были применены 2 характеристики:

1. Лексический спектр текста на уровне словаря (f).
2. Лексический спектр текста на уровне текста (mf).

Для определения лексического спектра, основанного на частотном словаре текста, используется следующий метод. Для сравнения текстов выбираются выборки одинаковой длины, например, 500 словоформ. Затем все словоформы распределяются по отдельным группам на основе частоты встречаемости m (1, 2, ..., 10 и более раз). После чего программа определяет количество словоформ в каждой группе f и покрываемость текста mf (для чего число словоформ в каждой группе умножается на частоту встречаемости слов из этой группы). Первый случай отображает распределение частот на уровне словаря, второй — на уровне текста. Эти характеристики можно проанализировать с помощью критерия Колмогорова-Смирнова для проверки однородности текста [Рогов 2021].

Таким образом, использование частотного словаря и анализ лексического спектра текста позволяют оценить распределение слов и их частоты в тексте и провести статистический анализ с целью определения характерных особенностей и однородности текста.

В качестве примера рассмотрим представленные в таблице 1 значения лексических спектров на уровне словаря (f) для первых фрагментов (500 слов) текстов под номерами 1 «Мелочи» Dubia («Время», 1861, № 6), 2 «Бесцветные явления. - Быль молодцу не укор. Новая комедия г. Н. Потехина. “Отеч. Записки”, №7 Июль» Dubia («Время», 1861, № 8), 5 «Наши домашние дела. (Современные заметки). Журнальные интересы» Dubia («Время», 1861, №9) и 7 «Записки князя Талейрана. (Собранные и изданные графиней О… дю К…, перевод с французского. Четыре части. Москва 1861 года)» Dubia («Время», 1862, №2). Значения лексических спектров были получены с помощью ИС «СМАЛТ» [Рогов 2019: 234-240].

Таблица 1

Фрагмент таблицы со значениями лексических спектров

m	Текст №1 (*f*)	Текст №3 (*f*)	Текст №5 (*f*)	Текст №7 (*f*)
1	207	225	195	207
2	49	33	41	32
3	11	11	14	16
4	6	6	10	5
5	4	7	3	2
6	2	1	4	4
7	1	2	0	3
8	0	2	0	2
9	2	1	3	4
10	6	4	5	4

На рисунке 1 изображена сравнительная диаграмма значений лексического спектра на уровне словаря (*f*) для текстов под номерами 1, 3, 5 и 7. Обратим внимание на то, что значение признаков уменьшается с увеличением параметра *m*.

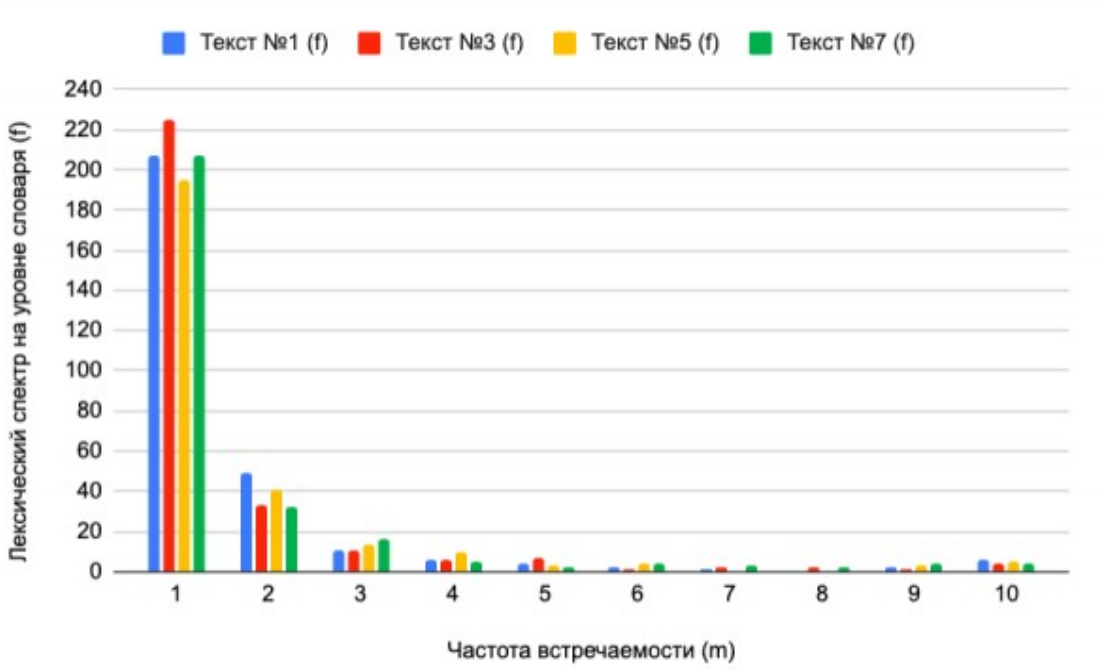

Рис. 1. Диаграмма распределения лексического спектра

Аппроксимация лексических спектров

Однако при решении задачи атрибуции текстов становится необходимым представление диаграммы одним числом при использовании многомерных методов, таких как деревья решений. Для этого можно аппроксимировать зависимость с помощью [Rogov: 223—229]:

1. Гиперболической кривой вида $y = \frac{a}{x} + b$.
2. Экспоненциальной кривой вида $y = c * e^{-\lambda x}$.
3. Степенной кривой вида $y = p * x^n$.

Помимо вышеперечисленных кривых, добавим в исследование логарифмическую кривую вида $y = s * \ln(x) + t$.

Для того, чтобы вычислить параметры кривых, необходимо предварительно нормировать исходный массив, содержащий значения лексических спектров. Параметры для построения кривых можно вычислить с помощью метода наименьших квадратов. В таблице 2 представлены примеры подсчета параметров кривых для аппроксимации лексического спектра на уровне словаря (*f*) для текстов под номерами 1, 3, 5 и 7.

Таблица 2

Примеры подсчета параметров кривых для аппроксимации лексического спектра

Параметры	Текст №1 (*f*)	Текст №3 (*f*)	Текст №5 (*f*)	Текст №7 (*f*)
a	77,456	81,055	75,727	77,942
b	-12,687	-13,741	-12,180	-12,829
c	27,216	24,389	30,662	20,087
λ	0,508	0,452	0,543	0,372
p	60,953	50,174	69,157	41,253
n	-2,384	-2,122	-2,514	-1,829
s	-23,332	-22,235	-26,485	-24,862
t	47,989	49,070	47,171	47,653

Отметим, что ранее аппроксимация гиперболической кривой распределения степеней графов фольклорных песен была применена в работе Н. Д. Москина [Москин]. Остановимся более подробно на логарифмической кривой. Для построения логарифмической кривой необходимо вычислить значения переменной *y* для всех значений *x*, подставив в выражение $y = s * \ln(x) + t$ значения характеристик *s* и *t*.

Для проверки гипотезы необходимо вычислить коэффициент корреляции Пирсона между модулем разности коэффициентов регрессии $d\frac{(1)}{ij} = |s_i - s_j|$ и расстоянием χ^2 между диаграммами (лексический спектр на уровне словаря (*f*)). Для этого обозначим z_{ij} значение спектра с номером *j* для текстового фрагмента *i* (*i = 1, 2, ..., 326)*, если $j \leq 9$, и сумма оставшихся значений, если $j = 10$. Расстояние вычисляется по формуле, где $z_i = \sum_{k=1}^{10} z_{ik}$ — сумма степеней по строкам (*i = 1, 2, ..., 326*), $z_i = \sum_{k=1}^{10} z_{jk}$— сумма степеней по столбцам (*j = 1, 2, ..., 326*). Матрицы расстояний χ^2 и модулей разности $d\frac{(1)}{ij} = |s_i - s_j|$ для лексического спектра на уровне словаря (*f*) для исходных текстов представлены в таблицах 3 и 4 соответственно.

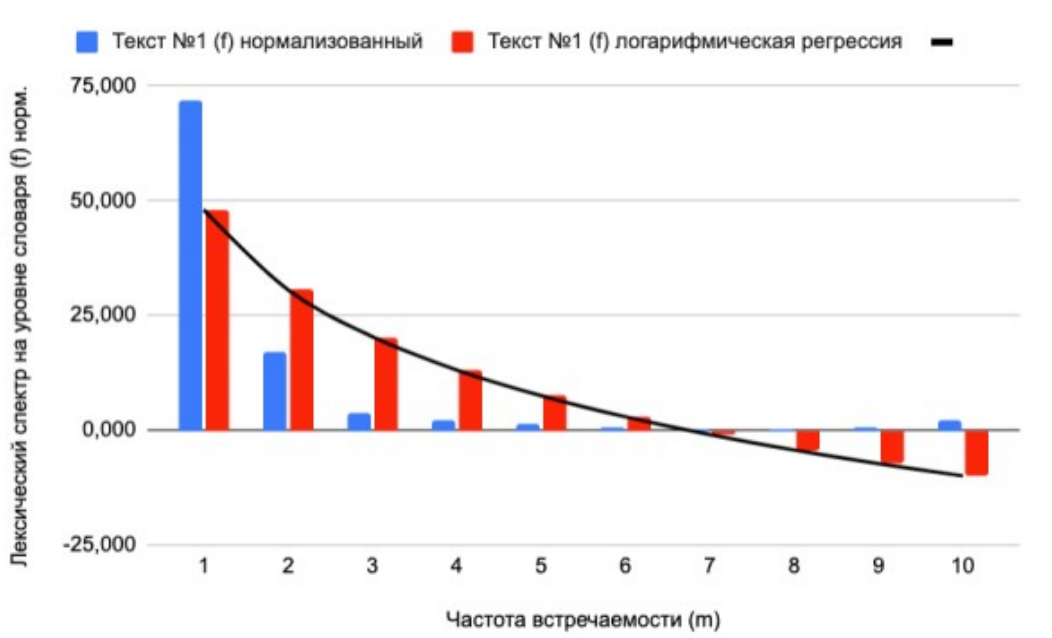

Рис. 2. Аппроксимация логарифмической кривой лексического спектра

Таблица 3

Матрица расстояний χ2 для лексического спектра

	1	2	3	4	5	6	7	…	326
1	0	7,131	11,592	4,508	8,263	9,262	5,917	…	2,434
2	7,131	0	2,515	3,594	6,080	8,352	8,001	…	4,036
3	11,592	2,515	0	5,997	8,158	8,010	14,951	…	6,181
4	4,508	3,594	5,997	0	2,318	3,907	5,319	…	3,457
5	8,263	6,080	8,158	2,318	0	4,469	7,335	…	6,082
6	9,262	8,352	8,010	3,907	4,469	0	6,388	…	6,598
7	5,917	8,001	14,951	5,319	7,335	6,388	0	…	6,515
…	…	…	…	…	…	…	…	…	…
326	2,434	4,036	6,181	3,457	6,082	6,598	6,515	…	0

Таблица 4

Матрица модулей разности ` $d\frac{(1)}{ij} = |s_i - s_j|$` для лексического спектра

	1	2	3	4	5	6	7	…	326
1	0	0,735	2,007	0,411	1,057	1,103	0,367	…	0,974
2	0,735	0	1,272	0,324	0,322	0,368	1,102	…	0,239
3	2,007	1,272	0	1,596	0,950	0,904	2,373	…	1,032
4	0,411	0,324	1,596	0	0,646	0,692	0,778	…	0,563
5	1,057	0,322	0,950	0,646	0	0,046	1,424	…	0,083
6	1,103	0,368	0,904	0,692	0,046	0	1,470	…	0,129
7	0,367	1,102	2,373	0,778	1,424	1,470	0	…	1,341
…	…	…	…	…	…	…	…	…	…
326	0,974	0,239	1,032	0,563	0,083	0,129	1,341	…	0

Коэффициенты корреляции Пирсона для первого параметра логарифмической регрессии s составили r = 0,6309436484, 0,6543479707, 0,6831338974, 0,6863642009` ` для текстовых фрагментов в количестве 100, 200,

300 и 326 текстов. По шкале Чеддока (таблица 5) связь считается «заметной». Обратим внимание на то, что теснота связи становится больше с увеличением количества текстовых фрагментов.

Таблица 5

Шкала Чеддока

Количественная мера тесноты связи	Качественная характеристика силы связи
0,1 – 0,3	Слабая
0,3 – 0,5	Умеренная
0,5 – 0,7	Заметная
0,7 – 0,9	Высокая
0,9 – 0,99	Весьма высокая

Дополнительно рассмотрим второй параметр логарифмической регрессии — t. Матрица модулей разности $d\frac{3}{ij} = |t_i - t_j|$ представлена в таблице 6. Коэффициенты корреляции Пирсона для второго параметра логарифмической регрессии t составили r = 0,6309436484, 0,6543479707, 0,6831338974, 0,6863642009 для текстовых фрагментов в количестве 100, 200, 300 и 326 текстов. Коэффициенты корреляции полностью совпали для первого и второго параметра логарифмической регрессии.

Таблица 6

Матрица модулей разности $d\frac{3}{ij} = |t_i - t_j|$ для лексического спектра

	1	2	3	4	5	6	7	…	326
1	0	0,735	2,007	0,411	1,057	1,103	0,367	…	0,974
2	0,735	0	1,272	0,324	0,322	0,368	1,102	…	0,239
3	2,007	1,272	0	1,596	0,950	0,904	2,373	…	1,032
4	0,411	0,324	1,596	0	0,646	0,692	0,778	…	0,563
5	1,057	0,322	0,950	0,646	0	0,046	1,424	…	0,083
6	1,103	0,368	0,904	0,692	0,046	0	1,470	…	0,129
7	0,367	1,102	2,373	0,778	1,424	1,470	0	…	1,341
…	…	…	…	…	…	…	…	…	…
326	0,974	0,239	1,032	0,563	0,083	0,129	1,341	…	0

Заключение. В ходе исследования была изучена теория, касающаяся математических

методов для решения задач атрибуции, в частности методика атрибуции Гейра Хьетсо, и анализа лексических спектров. Была аппроксимирована зависимость лексических спектров текстов с помощью логарифмической кривой и были вычислены коэффициенты корреляции Пирсона для параметров логарифмической регрессии для разного количества фрагментов (всего 326 текстов). Коэффициенты корреляции Пирсона для первого и второго параметров логарифмической регрессии полностью совпали. Для данных выборок теснота связи считается «заметной». Теснота связи становится больше с увеличением количества фрагментов в выборке. Это означает, что при использовании большего числа текстовых фрагментов для анализа статистические закономерности в стилистическом почерке и авторской манере писателя становятся более заметными.

СПИСОК ЛИТЕРАТУРЫ

Кулаков К. А., Лебедев А. А., Рогов А. А., Суровцова Т. Г., Москин Н. Д. Атрибуция текстов с помощью математических методов и компьютерных технологий // Материалы XIII всероссийской научно-практической конференции «Цифровые технологии в образовании, науке, обществе» (Петрозаводск, 17—20 сентября 2019 года). Петрозаводск, 2019. С. 121—125.

Москин Н. Д. Теоретико-графовые модели фольклорных текстов и методы их анализа. Петрозаводск: Изд-во ПетрГУ, 2013. 148 с.

Рогов А. А., Абрамов Р. В., Бучнева Д. Д., Захарова О. В., Кулаков К. А., Лебедев А. А., Москин Н. Д., Отливанчик А. В., Савинов Е. Д., Сидоров Ю. В. Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин» Петрозаводск : Издательство «Острова», 2021. 391 с.

Рогов А. А., Кулаков К. А., Москин Н. Д. Программная поддержка в решении задачи атрибуции текстов // Программная инженерия. 2019. Т. 10. № 5. – С. 234—240.

Хьетсо Г. Принадлежность Достоевскому: к вопросу об атрибуции Ф. М. Достоевскому анонимных статей в журналах «Время» и «Эпоха». Осло, 1986. 82 с.

Rogov A., Moskin N., Kulakov K., Abramov R. Machine Learning Methods in the Problem of Attribution of Publicistic Texts of the XIX Century // Proceedings of the 30th Conference of Open Innovations Association FRUCT. 2021. Vol. 30. P. 223—229.

Original article

Demid A. KUUSELA

bachelor’s degree, Petrozavodsk State University (Petrozavodsk, Russia),
demid.kuusela@gmail.com

ANALYSIS OF LEXICAL SPECTRA OF TEXTS USING MATHEMATICAL METHODS

Scientific adviser:
Nikolai D. Moskin
Paper submitted on: 06/09/2023;
Accepted on: 06/17/2023;
Published on: 06/25/2023

Abstract. The article deals with the analysis of lexical spectra of texts using mathematical methods and computer technologies. In particular, the approximation of the values of lexical spectra of texts using a logarithmic curve is described.

Keywords: text attribution, establishment of authorship, lexical spectrum, approximation, information system «SMALT»

For citation: Kuusela D. A. Analysis of lexical spectra of texts using mathematical methods. *StudArctic Forum*. 2023; 8(2): 30—35.