УДК 81-139+81-112 doi: 10.15827/0236-235X.142.654-660 2023. Т. 36. № 4. С. 654–660

Поиск оптимального набора букв для стилевой классификации художественных текстов методом статистических индексов

Л.Г. Горбич

Ссылка для цитирования
Горбич Л.Г. Поиск оптимального набора букв для стилевой классификации художественных текстов методом статистических индексов // Программные продукты и системы. 2023. Т. 36. № 4. С. 654–660. doi: 10.15827/0236-235X.142.654-660
Информация о статье
Поступила в редакцию: 30.06.2023 После доработки: 12.07.2023 Принята к публикации: 14.07.2023

Аннотация. В статье рассматривается проблема улучшения методов стилевой классификации русскоязычных текстов. В качестве возможного направления исследований предложен метод оптимизации набора (множества) букв, применяемого для вычисления статистических индексов текстов. Для оптимизации и контроля результатов использованы поэтические и прозаические художественные тексты на русском языке. Объем текстов составлял порядка 300 тысяч знаков при оптимизации и 100 тысяч знаков при контрольной оценке. Для вычисления статистических индексов рассчитывались частотности биграмм и триграмм букв. При оптимизации опробован также и вариант совместного использования индексов биграмм и триграмм. В статье дано краткое описание метода статистических индексов, приведены применявшиеся в исследовании алгоритм пошаговой оптимизации, вид возможной оптимизационной функции и формула для нахождения границы классификации. Показано, что оптимизация набора букв улучшает классификацию по сравнению с вариантом использования как полного набора букв, так и набора из гласных букв в применении к задаче автоматического различения поэтических и прозаических художественных текстов на русском языке. Проведено сравнение результатов классификации по предложенной формуле границы классификации с результатами расчетов по классификации методом ROC-кривых. В итоге для разных сочетаний статистических индексов и способов определения границы классификации интервал верной классификации составил 72–74 % для набора, включающего все буквы, 82–86 % для набора, включающего только гласные буквы, и 80.5–92.5 % для разных наборов букв, полученных при оптимизации.
Ключевые слова: стилевая классификация, набор букв, автоматическая классификация текстов, статистический индекс, машинное обучение, метод оптимизации, ROC-кривая

Введение. Методы исследования текстов на основе расчета частотностей букв и их сочетаний были предложены очень давно [1] и пережили всплеск интереса к ним в связи с развитием компьютерных исследований [2, 3]. Подходы, основанные на графемах, естественным образом дополняют лексические, такие как анализ частотностей отдельных слов и словосочетаний [4], расчеты встречаемости определенных частей речи [5, 6]. В качестве значимых показателей исследователями рассматривались и другие свойства текстов, например, средняя длина предложения, ритмические характеристики текста [7] или индекс сжатия стандартными компьютерными алгоритмами [8]. Эти методы применялись для решения таких задач, как авторская атрибуция, стилевая классификация, определение тональности (эмоциональной окрашенности) различных текстов.

Методы, основанные на частотностях n-грамм букв, успешно применяются и в настоящее время как для русского языка [9, 10], так и для широкого спектра других языков [11, 12]. Однако такие исследования носят, скорее, феноменологический характер, поскольку теоретически обосновать значимость какого-либо сочетания букв сложнее, чем объяснить использование в тексте самостоятельных или служебных частей речи. Соответственно, и улучшение таких методов видится как экспериментально-эмпирическая задача. На той же слабости теоретического обоснования, по-видимому, основана и тенденция рассмотрения в исследованиях полного набора букв и буквосочетаний, поскольку объяснить выбор каких-то конкретных букв трудно и, как следствие, для статистического анализа все они видятся равнозначимыми.

В работах [13, 14] для вычисления статистических индексов использован набор гласных букв вследствие изначального предположения, что эти буквы несут меньше осознаваемой информативной нагрузки по сравнению с согласными. Успешное применение такого подхода (как и улучшение характеристик распознавания по сравнению с использованием полного набора букв) позволяет сформулировать проблему: можно ли, оптимизируя набор букв, используемых в вычислениях, улучшить показатели классификации? Ответу на этот вопрос и посвящена данная работа.

Используемые тексты

Для нахождения оптимального набора букв использовались русскоязычные тексты 19 авторов, имеющих достаточные объемы произведений как в поэзии, так и в прозе. Для каждого автора подбирались по два текста: один поэтический и один прозаический, что позволило выровнять выборку и уменьшить влияние индивидуальных лексических особенностей. Поскольку отдельные произведения авторов могут быть малы и недостаточны для статистического анализа (особенно это характерно для поэзии), в качестве текстов при оптимизации использовались корпуса произведений авторов. Объем каждого сборного текста был 300 тысяч символов. Использовались сочинения К.Д. Бальмонта, В.Я. Брюсова, И.А. Бунина, Д.Л. Быкова, В.С. Высоцкого, Е.А. Евтушенко, Н.М. Карамзина, И.А. Крылова, М.Ю. Лермонтова, С.Я. Маршака, В.В. Набокова, Н.А. Некрасова, Б.Ш. Окуджавы, Б.Л. Пастернака, А.С. Пушкина, К.М. Симонова, Ф.К. Сологуба, А.К. Толстого, Д.И. Хармса.

Пороги классификации вычислялись по фрагменту из первых 100 тысяч символов текстов, используемых для оптимизации.

Оценка релевантности найденных параметров классификации проводилась по массиву из 100 сборных текстов поэтического подстиля и 100 текстов прозаического подстиля разных авторов. При этом для контроля использовались тексты тех авторов, которые не входят в приведенный выше список. Каждый сборный текст принадлежал только одному автору. Объем использованных для контроля текстов составлял около 100 тысяч знаков.

Расчет статистических индексов

В качестве факторов классификации были использованы статистические индексы n-грамм букв, которые по своей сути являются интегральными показателями скоррелированности последовательно идущих букв. Чем ближе вероятности обнаружения сочетаний букв в последовательности к вероятностям для независимых событий, тем меньше величина индекса. Отметим, что *индексы биграмм* (ИБ) и *триграмм* (ИТ) букв в исследовании рассчитывались следующим образом: из текста исключались все буквы, кроме входящих в набор; оставшиеся буквы рассматривались как единая строка; заглавные и строчные варианты написания учитывались как одинаковые буквы; буква «ё» учитывалась как «е»; рассчитывалось количество двоек (или троек) последовательно идущих букв для каждого возможного сочетания; рассчитывался индекс по формуле критерия согласия Пирсона χ^2:

$$\chi^2 = \sum_{i=1}^{k} \frac{(p_i^{theor} - p_i^{emp})^2}{p_i^{theor}},$$

где p_i^{theor} – вероятность обнаружения биграмм и триграмм, вычисляемая как произведение вероятностей обнаружения отдельных букв; p_i^{emp} – вероятность, полученная из обработки реального текста; k – количество возможных сочетаний.

Алгоритм поиска

Используемый алгоритм вычисления статистических индексов не позволяет найти оптимальный набор букв простым расчетом частотностей всех их сочетаний для полного текста и последующим отбором наиболее значимых комбинаций. При таком способе расчета индексов становятся значимыми и графемы, входящие в набор, и те, которые в набор не входят и при расчете индексов пропускаются, поскольку пропуск букв порождает новые сочетания графем, не присутствующие изначально в анализируемом тексте. В то же время полный перебор всех возможных наборов букв для нахождения оптимального потребовал бы очень большого расчетного времени. Поэтому подбор в исследовании проводился пошаговым методом по следующему алгоритму.

1. Случайным образом генерировался исходный набор букв, и для него вычислялось значение оптимизационной функции.
2. На основании исходного набора формировались наборы-кандидаты путем удаления или добавления одной буквы. Для каждого такого набора рассчитывалось значение оптимизационной функции.
3. Если обнаруживалось, что существуют наборы с большим, чем в исходном наборе, значением оптимизационной функции, то наиболее удачный набор принимался в качестве текущего и алгоритм повторялся с шага 2.
4. Если текущий набор оказывался самым оптимальным из сравниваемых вариантов, значит, был достигнут локальный максимум и этот вариант запоминался.
5. Поскольку локальных максимумов в пространстве оптимизации может быть много (что подтверждалось на практике), вычисления повторялись с шага 1.

6. После накопления достаточной статистики по локальным максимумам лучший набор букв принимался как итоговый.

В качестве оптимизационной функции в исследовании использовалось значение нормированного расстояния между центрами кластеров:

$$f = \frac{2 \cdot |\bar{x}_2 - \bar{x}_1|}{sd_2 + sd_1}, \quad (1)$$

где $\bar{x}$ – средние арифметические значения индексов для каждого кластера; sd – их среднеквадратические отклонения. При оптимизации осуществлялся поиск максимального значения этой функции. Отметим, что при максимизации функции такого вида не только кластеры точек раздвигаются относительно друг друга, но и уменьшается разброс точек внутри кластера.

Если оптимизация осуществлялась одновременно по двум индексам (ИБ и ИТ), то общее значение функции рассчитывалось как

$$f_{\Sigma} = \sqrt{f_{\text{ИБ}}^2 + f_{\text{ИТ}}^2}, \quad (2)$$

где $f_{\text{ИБ}}$ и $f_{\text{ИТ}}$ – значения оптимизационной функции, вычисленные отдельно для индексов ИБ и ИТ.

За границу классификации принималось значение, соответствующее равенству нормированного расстояния до центров кластеров:

$$x_{\text{Г}} = \frac{\bar{x}_2 \cdot sd_1 + \bar{x}_1 \cdot sd_2}{sd_1 + sd_2}. \quad (3)$$

Для сравнения методов использовался также и другой способ классификации – метод ROC-кривых. Его принципиальное отличие в том, что для нахождения порога классификации используются не те или иные выражения для расстояний, а количество правильно классифицированных точек в обучающей выборке.

Результаты расчетов и проверка классифицирующей способности по методу нормированных расстояний

Оценка качества результатов оптимизационных процедур проводилась путем сравнения матриц ошибок (confusion matrix) классификации текстов из контрольной выборки. Основой для сравнения стали результаты расчетов для неоптимизированных наборов букв. Рассчитанные матрицы ошибок для всех букв (абвгдежзийклмнопрстуфхцчшщъыьэюя) приведены в таблице 1, расчеты для набора из гласных букв (аеиоуыэюя) – в таблице 2. Граница классификации находилась по тем же 19 парам текстов и по той же формуле, что и для оптимизированных наборов букв.

Таблица 1

Результаты классификации по набору из всех букв для индексов биграмм и триграмм

Table 1

Classification results by all letters for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	84	16	Поэзия	79	21
Проза	40	60	Проза	32	68

Таблица 2

Результаты классификации по набору из гласных букв для индексов биграмм и триграмм

Table 2

Classification results by vowels for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	74	26	Поэзия	74	26
Проза	5	95	Проза	3	97

При оптимизации только по индексу биграмм лучшие результаты были получены для набора букв «абгелмошщыэя», значение оптимизационной функции составило 3.81. Результаты классификации 200 текстов для этого набора приведены в таблице 3.

Таблица 3

Результаты классификации по набору «абгелмошщыэя» для индексов биграмм и триграмм

Table 3

Classification results by «абгелмошщыэя» for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	98	2	Поэзия	95	5
Проза	17	83	Проза	12	88

При оптимизации только по индексу триграмм лучшие результаты были получены для набора букв «бгеичшы», значение оптимизационной функции равно 2.99. Результаты контрольной классификации для этого набора приведены в таблице 4.

При оптимизации по функции, учитывающей одновременно индексы биграмм и триграмм, лучшим набором оказался «абгеочш щыэя». Значение оптимизационной функции составило 3.68. Результаты классификации контрольных текстов приведены в таблице 5.

Таблица 4

Результаты классификации по набору «бгеичшы» для индексов биграмм и триграмм

Table 4

Classification results by «бгеичшы» for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	97	3	Поэзия	96	4
Проза	30	70	Проза	22	78

Таблица 5

Результаты классификации по набору «абгеочшщыэя» для индексов биграмм и триграмм

Table 5

Classification results by «абгеочшщыэя» for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	99	1	Поэзия	94	6
Проза	26	74	Проза	9	91

Пороговые значения при классификации, найденные по (3), для вычисленных индексов ИБ и ИТ для всех наборов букв приведены в таблице 6.

Таблица 6

Пороговые значения индексов при классификации по расстояниям для разных наборов букв

Table 6

Threshold values of indices when classifying by distances for different letter sets

Описание	Набор букв	ИБ	ИТ
Все буквы	абвгдежзийклмнопр стуфхцчшщъыьэюя	0.974	6.073
Гласные	аеиоуыэюя	0.022	0.073
Подбор по ИБ	абгелмошщыэя	0.131	0.418
Подбор по ИТ	бгеичшы	0.047	0.111
Подбор по ИБ и ИТ	абгеочшщыэя	0.097	0.270

Проверка классифицирующей способности при использовании метода ROC-кривых

Вторым методом нахождения порога классификации, применявшимся для оценки классифицирующей способности при использовании различных наборов букв, был метод ROC-кривых. Были использованы те же оптимизированные наборы букв и те же контрольные тексты, что и для метода классификации по нормированным расстояниям. Матрицы ошибок классификации, полученные при расчетах по этому методу, представлены в таблицах 7–11, а в таблице 12 приведены найденные пороговые значения.

Таблица 7

Результаты ROC-классификации по набору из всех букв для индексов биграмм и триграмм

Table 7

ROC classification results by all letters for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	84	16	Поэзия	66	34
Проза	36	64	Проза	22	78

Таблица 8

Результаты ROC-классификации по набору из гласных букв для индексов биграмм и триграмм

Table 8

ROC classification results by vowels for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	65	35	Поэзия	76	24
Проза	1	99	Проза	4	96

Таблица 9

Результаты ROC-классификации по набору «абгелмошщыэя» для индексов биграмм и триграмм

Table 9

ROC classification results using «абгелмошщыэя» for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	100	0	Поэзия	99	1
Проза	25	75	Проза	18	82

Таблица 10

Результаты ROC-классификации по набору «бгеичшы» для индексов биграмм и триграмм

Table 10

ROC classification results by «бгеичшы» for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	98	2	Поэзия	98	2
Проза	37	63	Проза	27	73

Таблица 11

Результаты ROC-классификации по набору «абгеочшщыэя» для индексов биграмм и триграмм

Table 11

ROC classification results by «абгеочшщыэя» for bigram and trigram indices

Индекс биграмм			Индекс триграмм		
	Прогноз			Прогноз	
Факт	Поэзия	Проза	Факт	Поэзия	Проза
Поэзия	99	1	Поэзия	95	5
Проза	26	74	Проза	17	83

Таблица 12

Пороговые значения индексов при ROC-классификации для разных наборов букв

Table 12

Threshold values of indices for ROC classification of different letter sets

Описание	Набор букв	ИБ	ИТ
Все буквы	абвгдежзийклмнопр стуфхцчшщъыьэюя	0.971	5.801
Гласные	аеиоуыэюя	0.021	0.074
Подбор по ИБ	абгелмошщыэя	0.137	0.440
Подбор по ИТ	бгеичшы	0.050	0.118
Подбор по ИБ и ИТ	абгеочшщыэя	0.097	0.278

Основные выводы

В работе ставились задачи определения целесообразности поиска оптимального набора графем и исследования работоспособности предложенного метода, а не получение максимально статистически обоснованного набора букв. Поэтому, хотя для подбора использовались лишь 19 пар текстов, подтверждение результативности метода на контрольной выборке в 200 текстов можно считать вполне убедительным.

Тем не менее асимметричность некоторых матриц ошибок позволяет надеяться, что с увеличением объема данных, используемых для оптимизации и вычисления пороговых значений, качество классификации еще больше возрастет.

Для удобства сравнения полученных результатов процент правильной классификации для различных наборов букв, двух индексов и двух способов классификации представлен в сводной таблице 13.

Результаты исследования позволили сделать следующие основные выводы.

Показатели классификации можно улучшить, найдя оптимальный набор букв. Предложенный метод пошаговой оптимизации применим для улучшения классификации текстов с помощью статистических индексов. Использованные в расчетах вид оптимизационной функции и формула для вычисления границы классификации показывают хорошую работоспособность. Процент правильной классификации в контрольной выборке возрастает по сравнению с классификацией как по всем буквам, так и по гласным буквам до 92.5.

Таблица 13

Процент правильной классификации для сочетания различных наборов букв, индексов и способов классификации

Table 13

Correct classification percentage for a combination of different sets of letters, indices and classification methods

Описание	Набор букв	ИБ, %	ИТ, %	ИБ, roc, %	ИТ, roc, %
Все буквы	абвгдежзийклмнопрстуфхцчшщъыьэюя	72.0	73.5	74.0	72.0
Гласные	аеиоуыэюя	84.5	85.5	82.0	86.0
Подбор по ИБ	абгелмошщыэя	90.5	91.5	87.5	90.5
Подбор по ИТ	бгеичшы	83.5	87.0	80.5	85.5
Подбор по ИБ и ИТ	абгеочшщыэя	86.5	92.5	86.5	89.0

Два способа нахождения пороговых значений классификации – метод, основанный на нормированных расстояниях, и метод ROC-кривых, показывают близкие результаты на контрольных выборках, однако метод ROC-кривых демонстрировал в среднем более слабые показатели качества классификации. При совместном использовании для оптимизации биграмм и триграмм результаты контрольной классификации оказались лучшими для индекса ИТ, но не самыми удачными для индекса ИБ.

Не всегда оптимальный набор, полученный для определенного индекса, являлся лучшим по результатам контрольных вычислений. Так, наборы букв, полученные при оптимизации по ИБ и ИБ+ИТ, демонстрируют лучшие результаты для индекса ИТ, чем набор, полученный при оптимизации собственно по ИТ.

В целом можно сделать заключение, что направление исследований, предложенное в работе, демонстрирует обнадеживающие результаты и может быть применимо при решении смежных задач классификации текстов.

Список литературы

1. Марков А.А. Примѣръ статистическаго изслѣдованія надъ текстомъ «Евгенія Онѣгина», иллюстрирующій связь испытаній въ цѣпь // Извѣстія Имп. Академіи Наукъ. 1913. Сер. VI. Т. 7. № 3. С. 153–162.
2. Хмелев Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестн. МГУ. Сер. 9: Филология. 2000. № 2. С. 115–126.
3. Keselj V., Peng F., Cercone N., Thomas C. N-gram-based author profiles for authorship attribution. Proc. PACLING, 2003, pp. 255–264.
4. Голубин Р.В., Судьин С.А., Дунаева Н.И., Афонин В.М., Ушаков А.В. Определение эмоциональной тональности текстов как инструмент социального управления: кейс COVID-19 // Теория и практика общественного развития. 2021. № 4. С. 13–19. doi: 10.24158/tipor.2021.4.1.
5. Kowsari K., Meimandi K., Heidarysafa M., Mendu S., Barnes L.E., Brown D.E. Text classification algorithms: A survey. Inf., 2019, vol. 10, no. 4, art. 150. doi: 10.3390/info10040150.
6. Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы. 2017. Т. 30. № 1. С. 85–99. doi: 10.15827/0236-235X.117.085-099.
7. Лагутина К.В., Лагутина Н.С., Бойчук Е.И. Классификация текстов по жанрам на основе ритмических характеристик // Моделирование и анализ информационных систем. 2021. Т. 28. № 3. С. 280–291. doi: 10.18255/1818-1015-2021-3-280-291.
8. Рябко Б.Я., Гуськов А.Е., Селиванова И.В. Теоретико-информационный метод классификации текстов // Проблемы передачи информации. 2017. Т. 53. № 3. С. 294–304.
9. Митин Н.А., Орлов Ю.Н. Статистический анализ биграмм специализированных текстов // Компьютерные исследования и моделирование. 2020. Т. 12. № 1. С. 243–254. doi: 10.20537/2076-7633-2020-12-1-243-254.
10. Воронина М.Ю., Кислицын А.А., Орлов Ю.Н. Алгоритм коррекции метода биграмм в задаче идентификации автора текста // Математическое моделирование. 2022. Т. 34. № 9. С. 3–20. doi: 10.20948/mm-2022-09-01.
11. Bojanowski P., Grave E., Joulin A., Mikolov T. Enriching word vectors with subword information. TACL, 2017, vol. 5, pp. 135–146. doi: 10.1162/tacl_a_00051.
12. Kruczek J., Kruczek P., Kuta M. Are n-gram categories helpful in text classification? Proc. ICCS. LNTCS, 2022, pp. 524–537. doi: 10.1007/978-3-030-50417-5_39.
13. Горбич Л.Г., Филимонов В.В., Живодёров А.А. Опыт различения поэтических и прозаических текстов на основе сравнения распределений биграмм гласных букв // Количественные методы в искусствознании: матер. Междунар. науч.-практич. конф. 2013. С. 163–166.
14. Горбич Л.Г., Живодеров А.А. Использование статистических индексов для различения научных и научно-популярных текстов на примере трудов А.Е. Ферсмана // Программные продукты и системы. 2020. Т. 33. № 4. С. 720–725. doi: 10.15827/0236-235X.132.720-725.

Software & Systems doi: 10.15827/0236-235X.142.654-660 2023, vol. 36, no. 4, pp. 654–660

Finding an optimal letter set for stylistic classification of literary texts by a statistical index method

Leonid G. Gorbich

For citation
Gorbich, L.G. (2023) 'Finding an optimal letter set for stylistic classification of literary texts by a statistical index method', *Software & Systems*, 36(4), pp. 654–660 (in Russ.). doi: 10.15827/0236-235X.142.654-660
Article info
Received: 30.06.2023 After revision: 12.07.2023 Accepted: 14.07.2023

Abstract. The paper concerns the problem of improving style classification methods of Russian-language texts. A method for optimizing a letter set for calculating statistical indices of texts is proposed as a possible research direction. The work used both poetic and prose literary Russian texts for optimization and monitoring the results. The volume of texts was about 300 thousand characters for optimization and about 100 thousand characters for control evaluation. Calculating statistical indices was based on calculating frequencies of letter bigrams and trigrams. Optimization also involved testing joint use of bigram and trigram indices. The paper provides a brief description of a method of statistical indices, a step-by-step optimization algorithm used in the study, a possible optimization function formula, and a formula for finding a classification boundary. It is shown that a letter set optimization improves classification in comparison with both using a full letter set

and using a vowel set, when applied to automatically distinguishing poetic and prose literary texts in Russian. Classification results are compared by the proposed classification boundary formula with ROC-curve method calculation results. Therefore, for different combinations of statistical indices and methods for determining a classification boundary, a correct classification range was: 72–74 % for an all letter set; 82–86 % for an only vowel set; 80.5–92.5 % for different sets of letters obtained during optimization.
Keywords: text style, automatic text classification, statistical index, machine learning, optimization method, ROC-curve

References

1. Markov, A.A. (1913) 'An example of a statistical study of the text "Eugene Onegin" illustrating the connection of tests in a chain', *News of the Imperial Academy of Sci. Ser. VI*, 7(3), pp. 153–162 (in Russ.).
2. Khmelev, D.V. (2000) 'Recognizing the author of the text using A.A. Markov chains', *Vestn. Moskovskogo Universiteta. Ser. 9. Philology*, (2), pp. 115–126 (in Russ.).
3. Keselj, V., Peng, F., Cercone, N., Thomas, C. (2003) 'N-gram-based author profiles for authorship attribution', *Proc. PACLING*, pp. 255–264.
4. Golubin, R.V., Sudin, S.A., Dunaieva, N.I., Afonin, V.M., Ushakov, A.V. (2021) 'The identifying the emotional content of texts as a social management tool: The case of COVID-19', *Theory and Practice of Social Development*, (4), pp. 13–19 (in Russ.). doi: 10.24158/tipor.2021.4.1.
5. Kowsari, K., Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L.E., Brown, D.E. (2019) 'Text classification algorithms: A survey', *Inf.*, 10(4), art. 150. doi: 10.3390/info10040150.
6. Batura, T.V. (2017) 'Automatic text classification methods', *Software & Systems*, 30(1), pp. 85–99 (in Russ.). doi: 10.15827/0236-235X.117.085-099.
7. Lagutina, K.V., Lagutina, N.S., Boychuk, E.I. (2021) 'Text classification by genre based on rhythm features', *Modeling and Analysis of Inform. Sys.*, 28(3), pp. 280–291 (in Russ.). doi: 10.18255/1818-1015-2021-3-280-291.
8. Ryabko, B.Ya., Gus'kov, A.E., Selivanova, I.V. (2017) 'Information-Theoretic method for classification of texts', *Problems of Inf. Transm.*, 53(3), pp. 294–304 (in Russ.).
9. Mitin, N.A., Orlov, Yu.N. (2020) 'Statistical analysis of bigrams of specialized texts', *Computer Research and Modeling*, 12(1), pp. 243–254 (in Russ.). doi: 10.20537/2076-7633-2020-12-1-243-254.
10. Voronina, M.Yu., Kislitsyn, A.A., Orlov, Yu.N. (2022) 'Algorithm of the correction of bigram method for the problem of the text author identification', *Math. Models and Comput. Simulations*, 34(9), pp. 3–20 (in Russ.). doi: 10.20948/mm-2022-09-01.
11. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T. (2017) 'Enriching word vectors with subword information', *TACL*, 5, pp. 135–146. doi: 10.1162/tacl_a_00051.
12. Kruczek, J., Kruczek, P., Kuta, M. (2022) 'Are n-gram categories helpful in text classification?', *Proc. ICCS. LNTCS*, pp. 524–537. doi: 10.1007/978-3-030-50417-5_39.
13. Gorbich, L.G., Filimonov, V.V., Zhivodyorov, A.A. (2013) 'The experience of distinguishing poetic and prose texts based on comparing the distributions of vowel bigrams', *Proc. Quantitative Methods in Art Studies*, pp. 163–166 (in Russ.).
14. Gorbich, L.G., Zhivoderov, A.A. (2020) 'Using statistical indexes to distinguish between scientific and popular science texts on the example of the works of A. E. Fersman', *Software & Systems*, 33(4), pp. 720–725 (in Russ.). doi: 10.15827/0236-235X.132.720-725.

Авторы
Горбич Леонид Геннадьевич [1],
научный сотрудник, glg@cbibl.uran.ru

[1] Центральная научная библиотека УрО РАН,
г. Екатеринбург,
620137, Россия

Authors
Leonid G. Gorbich [1],
Research Associate, glg@cbibl.uran.ru

[1] Central Scientific Library
of the Ural Branch of the RAS,
Yekaterinburg, 620137, Russian Federation