ИЗДАТЕЛЬСТВО грамота

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

ISSN 2782-4543 (online)
ISSN 1997-2911 (print)

2023. Том 16. Выпуск 11 | 2023. Volume 16. Issue 11
Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Расширение стандартного сбалансированного лингвистического корпуса, построенного по правилам spaCy, коннотативными характеристиками

Горожанов А. И.

Аннотация. Ставится цель разработать технологию автоматического определения тональности текста на базе имеющегося авторского программного комплекса. Научная новизна заключается в том, что в работе предлагается структурно-функциональная модель полностью автоматизированного процесса оценки тональности текста в совокупности с анализом его морфологических характеристик; также впервые вводятся технические термины «коннотативная амплитуда» и «коннотативная плотность». В ходе исследования была построена модель базы данных, которая вмещает коннотативные числовые параметры; далее, написан программный код «надстройки» генератора, которая позволяет дополнять стандартную базу данных этими параметрами; наконец, проведена апробация технологии на материале трех романов Ф. Кафки («Замок», «Процесс» и «Америка») и двух романов Э. М. Ремарка («На Западном фронте без перемен» и «Возлюби ближнего своего») на немецком языке. В результате доказывается, что «надстройка» является качественным программным продуктом, который не дает технических сбоев и способен предоставлять исследователю целый набор коннотативных данных для последующей комплексной интерпретации текста при условии качественного входного тонального словаря.

EN

Extension of a standard balanced linguistic corpus built according to spaCy rules by connotative characteristics

Gorozhanov A. I.

Abstract. The aim of the research is to develop the technology for automatically determining the sentiment of a text based on the existing author's software package. The scientific novelty lies in the fact that the work proposes a structural and functional model of a fully automated process for assessing the sentiment of a text in conjunction with an analysis of its morphological characteristics; the technical terms "connotative amplitude" and "connotative density" are also introduced for the first time. The study built a database model that accommodates connotative numeric parameters; further, the program code for the "add-on" for the database generator has been written, which allows one to supplement the standard database with these parameters; finally, the technology was tested on the material of three novels by F. Kafka ("Castle", "The Trial" and "America") and two novels by E. M. Remarque ("All Quiet on the Western Front" and "Flotsam") in the German language. As a result, it is proven that the "add-on" is a high-quality software product that does not cause technical failures and is capable of providing researchers with a whole set of connotative data for subsequent comprehensive interpretation of the text, on condition that the input tone dictionary is of high quality.

Введение

Проблема определения тональности текста и область прикладных (корпусных) лингвистических исследований тесно связаны между собой, о чем свидетельствует ряд актуальных публикаций (Глушак, 2023; Пронина, Пронин, 2023; Рудаковский, 2023).

Среди современных инструментов определения тональности текстов (сентимент-анализа) исследователи отмечают различные NLP-библиотеки (Алтышева, 2023, с. 53), к которым подключаются тональные словари для того или иного языка (Гончаров, Лысенкова, Назин, 2023); в последнее время все чаще задействуют нейросети (Семенова, 2022, с. 84).

Научная статья (original research article) | https://doi.org/10.30853/phil20230594

В настоящей работе предлагается авторский метод, который объединяет в себе применение технологий NLP и тональных словарей и является техническим продолжением разрабатываемого в лаборатории фундаментальных и прикладных проблем виртуального образования Московского государственного лингвистического университета программного комплекса – корпусного менеджера и генератора баз данных, которые мы условно назовем *стандартным сбалансированным лингвистическим корпусом*, имеющим автоматическую морфологическую разметку по правилам NLP-библиотеки spaCy (Горожанов, Степанова, 2022). Таким образом, описываемая технология расширяет созданное ранее, добавляя возможность включения в поисковые запросы коннотативных параметров. Указанные выше работы формируют теоретико-методологическую базу нашего прикладного исследования.

В пользу актуальности темы говорит значительное количество современных публикаций по проблеме определения тональности (сентимент-анализа) текста (Комарова, 2023; Груздева, Юрьев, Бессмертный, 2022; Панфилова, Ушаков, 2022).

Отдельного внимания заслуживает тот факт, что сентимент-анализ активно применяется сегодня в таких областях, как *искусственный интеллект* (Раббимов, 2022) и *информационная и информационно-психологическая безопасность* (Логинова, 2022), поэтому совершенствование технологий определения тональности текстов различных жанров является важным и актуальным делом не только с точки зрения науки, но также и для сферы народного хозяйства.

Мы ставим перед собой следующие задачи:

1. Расширить модель базы данных коннотативными параметрами.
2. Модифицировать код программного комплекса сбалансированного лингвистического корпуса для получения возможности выполнения новых поисковых запросов.
3. Провести апробацию созданной технологии на лингвистическом материале.

Лингвистическим материалом исследования являются оригинальные тексты трех романов Ф. Кафки («Замок», «Процесс» и «Америка») и двух романов Э. М. Ремарка («На Западном фронте без перемен» и «Возлюби ближнего своего»). В качестве технического материала, или инструментария, обозначим язык программирования Python, графическую библиотеку PyQt5, базу данных SQLite, а также банк данных SentiWS – тональный словарь немецкого языка, представленный в виде файла CSV и использованный нами частично.

В ходе решения первой задачи (расширения модели базы данных коннотативными параметрами) применяется метод моделирования; вторая задача (модифицирования кода программного комплекса) предполагает использование метода объектно-ориентированного программирования; на стадии решения третьей задачи (апробации полученной технологии) привлекается метод анализа.

Практическая значимость работы заключается в том, что в результате программный комплекс дополняется «надстройкой» для определения тональности текста лингвистического корпуса. Кроме того, полученная модель базы данных и программное обеспечение могут быть использованы при чтении таких учебных дисциплин, как «Корпусная лингвистика», «Интерпретация текста», «Профессионально ориентированное программирование» и т. п.

Обсуждение и результаты

Первая задача исследования заключалась в корректировании имеющейся модели базы данных сбалансированного лингвистического корпуса, которая в стандартном варианте состоит из двух таблиц (для предложений и для токенов), которые, в свою очередь, имеют по пять пустых колонок формата TEXT, зарезервированных для возможной дополнительной разметки (Горожанов, 2022, с. 3383-3384).

Новая модель таблицы токенов предполагает заполнение первой резервной колонки *tokenoption01*, которая при генерации корпуса равна *NONE* (см. Рисунок 1).

	Name	Data type	Primary Key	Foreign Key	Unique	Check	Not NULL	Collate	
1	id	integer							*NULL*
2	tokennum	integer							*NULL*
3	sent_num	integer							*NULL*
4	tokentext	text							*NULL*
5	tokenpos	text							*NULL*
6	tokenlemma	text							*NULL*
7	tokenattr	text							*NULL*
8	tokenoption01	text							'NONE'
9	tokenoption02	text							'NONE'
10	tokenoption03	text							'NONE'
11	tokenoption04	text							'NONE'
12	tokenoption05	text							'NONE'

Рисунок 1. *Структура таблицы токенов с указанием задействованной резервной колонки*

Заполнение происходит положительными, нулевыми или отрицательными значениями, соответствующими леммам таблицы в банке данных SentiWS, например *0,004* или *0,0*, или *-0,8*.

Если банк данных не содержит текущую лемму таблицы, то ячейка остается без изменений. Таким образом, в новой модели колонка *tokenoption01* предусматривает следующие значения: *NONE*, *отрицательная десятичная дробь*, *ноль* или *положительная десятичная дробь*.

Далее было необходимо определить характер данных для наполнения резервных колонок таблицы предложений.

Здесь нами была выдвинута гипотеза, что полезными для интерпретации текста могут быть следующие пять параметров:

1. Общее количество положительно и отрицательно коннотированных токенов в предложении (целое число).
2. Сумма положительно коннотированных токенов в предложении (целое число).
3. Сумма отрицательно коннотированных токенов в предложении (целое число).
4. Сумма значений всех коннотированных токенов в предложении (десятичная дробь).
5. Сумма значений по модулю всех коннотированных токенов в предложении (десятичная дробь).

Эти пять параметров должны занять все пять резервных колонок таблицы (см. Рисунок 2).

	Name	Data type	Primary Key	Foreign Key	Unique	Check	Not NULL	Collate	
1	id	integer	🔑						*NULL*
2	sentnum	integer					◉		*NULL*
3	senttext	text					◉		*NULL*
4	sentoption01	text							'NONE'
5	sentoption02	text							'NONE'
6	sentoption03	text							'NONE'
7	sentoption04	text							'NONE'
8	sentoption05	text							'NONE'

Рисунок 2. *Структура таблицы предложений с указанием пяти задействованных резервных колонок*

Общее количество положительно и отрицательно коннотированных токенов в предложении назовем *коннотативной плотностью* (ρ_{con}) предложения, а сумму значений по модулю всех коннотированных токенов в предложении – *коннотативной амплитудой* (A_{con}) предложения.

Указанные параметры были выделены исходя из интерпретации текста читателем и достаточно полно описывают тональность предложения. Так, коннотативная плотность предназначается для маркирования тонально окрашенных предложений, а коннотативная амплитуда показывает диапазон тональных значений, даже если в простой сумме они дадут ноль. В самом деле, если в предложении присутствуют четыре окрашенных токена со значениями *0,5*, *0,25*, *-0,5* и *-0,25*, то простая сумма значений будет равна нулю и формально предложение ничем не будет отличаться от такого, в котором нет ни одного окрашенного токена. При этом коннотативная амплитуда покажет значение:

$$A_{con} = 0,5 + 0,25 + |-0,5| + |-0,25| = 1,5.$$

Решение второй задачи заключалось непосредственно в написании программного кода, точнее – в модификации уже имеющегося. Здесь изменения коснулись только программы-генератора лингвистического корпуса, которая получила «надстройку» в виде трех дополнительных функций: а) для добавления значений в таблицу токенов, б) для добавления значений в таблицу предложений и в) для построения файлов CSV с целью последующей визуализации коннотативных характеристик сбалансированного корпуса. Код корпусного менеджера не претерпел изменений, что доказывает высокую степень его универсальности.

Заметим, что генерация стандартной базы данных строится по пути от таблицы предложений к таблице токенов, а ее расширение коннотациями, наоборот, – от таблицы токенов к таблице предложений.

В функции для добавления данных в таблицу токенов частично использовался алгоритм, который применяется в корпусном менеджере для генерации частотного списка по частям речи, что соответствует принципу *многократного использования* (англ. reusability). Поскольку в банке данных SentiWS присутствуют только существительные, прилагательные и глаголы, то из базы данных сбалансированного корпуса отбираются уникальные леммы именно по этим частям речи, затем происходит их последовательное перебирание с проверкой на условие совпадения с содержанием банка данных. При наличии совпадения в запись базы данных соответствующей леммы добавляется числовое коннотативное значение (см. Рисунок 3).

На Рисунке 3 представлен фрагмент таблицы токенов базы данных сбалансированного корпуса романа Ф. Кафки «Замок», в которой в строках с идентификатором 33 и 37 программа зафиксировала наличие коннотированных лемм. Ячейки других лемм остались без изменений.

Работа с таблицей предложений была более сложной, поскольку предполагала добавление сразу пяти параметров, расчет которых производился из данных единственной новой колонки таблицы токенов.

Для каждого предложения перебираются его токены, накапливаются общее количество и отдельно количество положительно и отрицательно коннотированных единиц, считается простая сумма значений и сумма по модулю.

28	27	3 ihn	PRON	ich	Case=Acc|Gender=Masc|Number=Sing|Per...	NULL
29	28	3 ,	PUNCT	,		NULL
30	29	3 auch	ADV	auch		NULL
31	30	3 nicht	PART	nicht		NULL
32	31	3 der	DET	der	Case=Nom|Definite=Def|Gender=Masc|Nu...	NULL
33	32	3 schwächste	ADJ	schwach	Case=Nom|Degree=Sup|Gender=Masc|Nu...	-0.9206
34	33	3 Lichtschein	NOUN	Lichtschein	Case=Nom|Gender=Masc|Number=Sing	NULL
35	34	3 deutete	VERB	deuten	Mood=Ind|Number=Sing|Person=3|Tense...	NULL
36	35	3 das	DET	der	Case=Nom|Definite=Def|Gender=Neut|Nu...	NULL
37	36	3 große	ADJ	groß	Case=Acc|Degree=Pos|Gender=Neut|Num...	0.3694
38	37	3 Schloß	NOUN	Schloß	Case=Acc|Gender=Neut|Number=Sing	NULL
39	38	3 an	ADP	an		NULL
40	39	3 .	PUNCT	.		NULL
41	40	4 Lange	ADV	langen		NULL
42	41	4 stand	VERB	stehen	Mood=Ind|Number=Sing|Person=3|Tense...	NULL

Рисунок 3. *Коннотативные значения по леммам schwach (слабый) и groß (большой)*

Эти данные записываются, соответственно, в пять дополнительных ячеек текущего предложения. Если в предложении не встречаются коннотированные токены, то ячейки заполняются нолями.

На Рисунке 4 приведен фрагмент таблицы предложений с уже заполненными дополнительными колонками для того же сбалансированного корпуса.

id	sentnum	senttext	sentoption01	sentoption02	sentoption03	sentoption04	sentoption05
1	1	Es war spät abends, als K ankam.	0	0	0	0.0	0.0
2	2	Das Dorf lag in tiefem Schnee.	0	0	0	0.0	0.0
3	3	Vom Schloßberg war nichts zu sehen, Neb...	2	1	1	-0.5512	1.29
4	4	Lange stand K auf der Holzbrücke, die von ...	1	1	0	0.004	0.004
5	5	Dann ging er, ein Nachtlager suchen;	0	0	0	0.0	0.0
6	6	im Wirtshaus war man noch wach, der Wirt...	1	0	1	-0.0474	0.0474
7	7	K war damit einverstanden.	0	0	0	0.0	0.0
8	8	Einige Bauern waren noch beim Bier, aber ...	1	1	0	0.0911	0.0911
9	9	Warm war es, die Bauern waren still, ein we...	1	0	1	-0.0048	0.0048
10	10	Aber kurze Zeit darauf wurde er schon gew...	0	0	0	0.0	0.0
11	11	Ein junger Mann, städtisch angezogen, mit...	2	2	0	0.008	0.008
12	12	Die Bauern waren auch noch da, einige hat...	1	1	0	0.3716	0.3716
13	13	Der junge Mensch entschuldigte sich sehr ...	2	2	0	0.008	0.008
14	14	Niemand darf das ohne gräfliche Erlaubnis.	1	1	0	0.004	0.004
15	15	Sie aber haben eine solche Erlaubnis nicht ...	1	1	0	0.004	0.004
16	16	K hatte sich halb aufgerichtet, hatte die Ha...	1	0	1	-0.0498	0.0498
17	17	Ist denn hier ein Schloß?"	0	0	0	0.0	0.0
18	18	"Allerdings", sagte der junge Mann langsa...	1	0	1	-0.0167	0.0167
19	19	"Und man muß die Erlaubnis zum Übernac...	2	2	0	0.3436	0.3436

Рисунок 4. *Заполненные дополнительные колонки таблицы предложений*

Третья функция «надстройки» программы-генератора отвечает за формирование трех файлов CSV на основе данных таблицы предложений. В первый файл записываются значения амплитуды для каждого предложения, во второй – значения плотности, а в третий – простая сумма. На основе этих данных можно легко построить графики коннотативной характеристики текста, где по оси X будут даны номера предложений, а по оси Y – одна из трех указанных характеристик.

Третья задача заключалась в апробации созданной технологии на лингвистическом материале. Для этого коннотативными значениями были расширены четыре стандартных сбалансированных лингвистических корпуса оригинальных текстов трех романов Ф. Кафки («Замок», «Процесс» и «Америка») и двух романов Э. М. Ремарка («На Западном фронте без перемен» и «Возлюби ближнего своего»), а также построены графики их коннотативных характеристик.

Приведем общие показатели тональности текстов для упомянутых романов. В первую очередь получим количество предложений с общей положительной, отрицательной и нейтральной коннотацией. Для этого троекратно воспользуемся опцией ручного запроса со следующими параметрами:

SELECT COUNT (*) FROM sents WHERE sentoption04 > 0.0
SELECT COUNT (*) FROM sents WHERE sentoption04 < 0.0
SELECT COUNT (*) FROM sents WHERE sentoption04 = 0.0

Результаты приведем в табличной форме (см. Таблицу 1).

Таблица 1. *Общие параметры тональности текстов*

Роман	Положительные	Нейтральные	Отрицательные	Всего предложений
«Замок»	1617	2195	1592	5404
«Процесс»	1111	1548	1094	3753
«Америка»	1334	1659	1032	4025
«На Западном фронте без перемен»	949	3106	1112	5167
«Возлюби ближнего своего»	2243	9963	1862	14068

Интересны также данные о максимальных величинах коннотативной амплитуды и плотности (см. Таблицу 2).

Таблица 2. *Максимальная амплитуда и плотность*

Роман	$A_{con\ (max)}$	$\rho_{con\ (max)}$
«Замок»	4,083	17
«Процесс»	3,5844	10
«Америка»	2,8019	11
«На Западном фронте без перемен»	2,3375	9
«Возлюби ближнего своего»	2,4204	10

Наконец, приведем график амплитуды для романа «Замок» и график простой суммы значений для романа «Возлюби ближнего своего» (см. Рисунки 5 и 6).

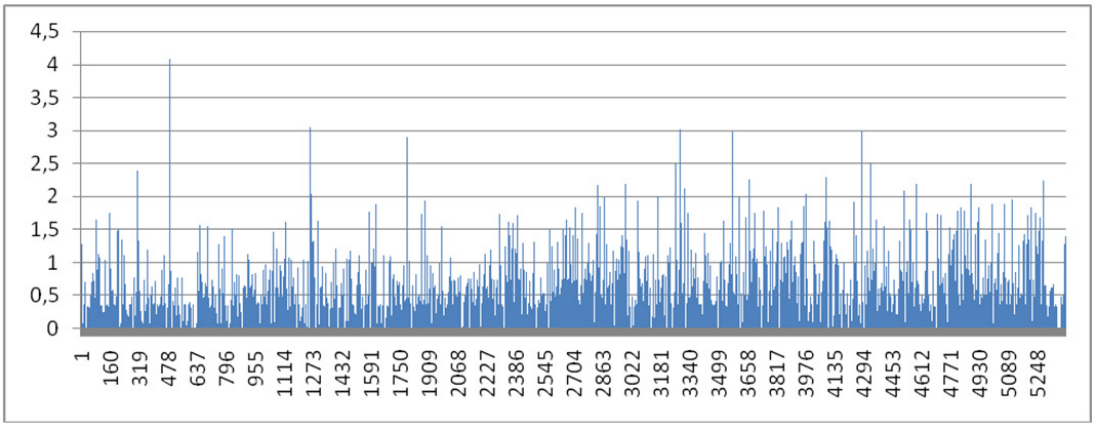

Рисунок 5. *График амплитуды для романа «Замок»*

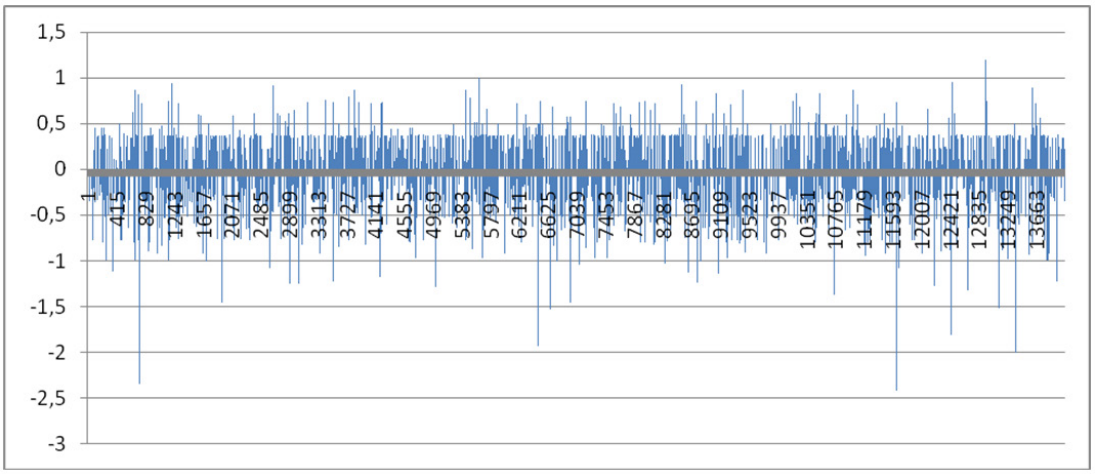

Рисунок 6. *График простой суммы значений для романа «Возлюби ближнего своего»*

Как было указано выше, по абсциссе здесь даны номера предложений, а по ординате – значение параметра.

Мы сознательно не анализируем полученные результаты с точки зрения литературоведения или лингвистической интерпретации текста, поскольку это выходит за рамки настоящей работы. Тем не менее мы полагаем, что созданная технология предоставляет лингвисту, филологу и литературоведу богатый материал для комплексных исследований художественного произведения.

Заключение

Итак, поставленная в рамках исследования цель была достигнута. Прежде всего, модель таблицы токенов базы данных сбалансированного лингвистического корпуса была дополнена колонкой числовых коннотативных характеристик, на основе которых удалось выделить пять новых параметров для таблицы предложений, в том числе *коннотативную плотность* и *коннотативную амплитуду*.

Далее, в файл генератора корпуса были внесены программные модификации, а именно – написаны три новые функции, отвечающие за заполнение таблиц токенов и предложений, а также за формирование специальных файлов CSV с коннотированными характеристиками корпуса.

Наконец, была осуществлена апробация созданного решения, в ходе которой в качестве лингвистического материала были привлечены пять произведений художественной литературы (два романа Э. М. Ремарка и три романа Ф. Кафки).

Однако нельзя не упомянуть, что представленная технология имеет ряд ограничений, связанных с качеством внешнего банка данных – тонального словаря. От его полноты и точности напрямую зависит достоверность интерпретации результата. В этой связи исследователю предлагается или использовать надежные банки данных, или составлять свои, которые могут представлять коннотативную характеристику не только в виде десятичной дроби, но также и в виде одного из компонентов трехчленной оппозиции («+», «0» или «–»). Созданная нами технология, при незначительных модификациях, допускает применение обоих вариантов.

Составление собственных банков данных – тональных словарей, в том числе и для различных языков, представляется нам одной из перспектив исследования.

Источники | References

1. Алтышева М. А. Проблемы и методы анализа русскоязычных текстов на предмет идентификации тональности // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2023. № 3.
2. Глушак В. М. Отрицание немецких полярных слов и выражений в автоматизированном анализе тональности текста // Филологические науки. Вопросы теории и практики. 2023. Т. 16. Вып. 10. https://doi.org/10.30853/phil20230510
3. Гончаров А. Р., Лысенкова С. А., Назин А. С. Формирование синонимичных рядов с экспертной оценкой для получения коэффициентов эмоциональности слов // Успехи кибернетики. 2023. Т. 4. № 2. https://doi.org/10.51790/2712-9942-2023-4-2-06
4. Горожанов А. И. Экспериментальное моделирование базы данных сбалансированного лингвистического корпуса // Филологические науки. Вопросы теории и практики. 2022. Т. 15. Вып. 10. https://doi.org/10.30853/phil20220563
5. Горожанов А. И., Степанова Д. В. Составление сбалансированного корпуса художественного произведения (на материале романов Ф. Кафки) // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2022. № 7 (862). https://doi.org/10.52070/2542-2197_2022_7_862_31
6. Груздева А. С., Юрьев Р. Н., Бессмертный И. А. Применение волновой модели текста к задаче сентимент-анализа // Научно-технический вестник информационных технологий, механики и оптики. 2022. Т. 22. № 6. https://doi.org/10.17586/2226-1494-2022-22-6-1159-1165
7. Комарова Е. В. Проблема цифрового этикета в русских и английских медиатекстах: на материале миграционного дискурса // Медиалингвистика. 2023. Т. 10. № 2. https://doi.org/10.21638/spbu22.2023.207
8. Логинова А. О. Подходы к обнаружению социальных интернет-ботов // Информация и безопасность. 2022. Т. 25. № 2. https://doi.org/10.36622/VSTU.2022.25.2.005
9. Панфилова А. С., Ушаков Д. В. Эмоциональный тон российского, итальянского, немецкого и французского новостного интернет-контента в период разворачивания пандемии COVID-19 // Психология. Журнал Высшей школы экономики. 2022. Т. 19. № 3. https://doi.org/10.17323/1813-8918-2022-3-562-586
10. Пронина Е. В., Пронин Д. Д. Исследовательский потенциал изучения корпуса произведений русской литературы с помощью цифровых лингвистических методов и технологий искусственного интеллекта (проект Lensky) // Современный ученый. 2023. № 3.
11. Раббимов И. М. Алгоритм построения ансамбля деревьев решений для сентиментального анализа текста // Проблемы вычислительной и прикладной математики. 2022. № 6 (45).
12. Рудаковский Я. С. Анализ тональности решений по денежно-кредитной политике Национального банка Республики Беларусь с помощью методов машинного обучения // Белорусский экономический журнал. 2023. № 3 (104). https://doi.org/10.46782/1818-4510-2023-3-115-126
13. Семенова М. О. Подходы к сентимент-анализу // Вестник Московского государственного лингвистического университета. Гуманитарные науки. 2022. № 12 (867). https://doi.org/10.52070/2542-2197_2022_12_867_83

Информация об авторах | Author information

RU

Горожанов Алексей Иванович[1], д. филол. н., доц.
[1] Московский государственный лингвистический университет

EN

Gorozhanov Alexey Ivanovich[1], Dr
[1] Moscow State Linguistic University

[1] *a_gorozhanov@mail.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 11.10.2023; опубликовано online (published online): 16.11.2023.

Ключевые слова (keywords): корпусная лингвистика; сбалансированный корпус; тональность текста; коннотация; немецкий язык; corpus linguistics; balanced corpus; sentiment of a text; connotation; German language.