УДК 004.912+801.6
DOI: 10.25559/SITITO.14.201804.888-895

ИЗВЛЕЧЕНИЕ ЛЕКСИЧЕСКИХ И МЕТРОРИТМИЧЕСКИХ ПРИЗНАКОВ, ХАРАКТЕРНЫХ ДЛЯ ЖАНРА И СТИЛЯ И ИХ КОМБИНАЦИЙ В ПРОЦЕССЕ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

В.Б. Барахнин[1,2], О.Ю. Кожемякина[1], Е.В. Рычкова[1,2], И.С. Пастушков[1], Ю.С. Борзилова[1]
[1] Институт вычислительных технологий Сибирского отделения Российской академии наук, г. Новосибирск, Россия
[2] Новосибирский национальный исследовательский государственный университет, г. Новосибирск, Россия

THE EXTRACTION OF LEXICAL AND METRORHYTHMIC FEATURES WHICH ARE CHARACTERISTIC FOR THE GENRE AND THE STYLE AND FOR THEIR COMBINATIONS WITHIN THE PROCESS OF AUTOMATED PROCESSING OF TEXTS IN RUSSIAN

Vladimir B. Barakhnin[1,2], Olga Yu. Kozhemyakina[1], Elena V. Rychkova[1,2], Ilya S. Pastushkov[1], Yuliya S. Borzilova[1]
[1] Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia

Ключевые слова

Распознавание образов; анализ главных компонент; автоматизированный анализ поэтических текстов; алгоритм классификации; ансамблирование.

Аннотация

В работе описан алгоритм извлечения характерных признаков для жанра и стиля. Работа выполнялась в рамках разработки программной системы, созданной в Институте вычислительных технологий СО РАН и предназначенной для комплексного анализа метроритмических и жанрово-стилистических характеристик поэтических текстов на русском языке. Система органично сочетает в себе как оригинальные программные модули, созданные непосредственно разработчиками системы и предназначенные для решения узкоспециализированных задач анализа поэтических текстов, так и программные продукты открытого доступа. Обобщённый подход, позволяющий рассматривать поэтические признаки в виде вектора, с одной стороны, позволяет использовать современные алгоритмы классификации и их ансамбли, с другой, такой подход имеет недостатки при малых объёмах корпусов, с которыми приходится работать. Поэтому наличие такого шага как верификация позволяет специалистам корректировать работу системы исходя из экспертных знаний, а также делает процесс классификации прозрачным. В качестве инструмента были применены библиотеки языка Python: scikit-learn, в которой реализованы алгоритмы классификации, а также методы их комбинирования, и ELI5, позволяющая установить соответствие между компонентами вектора признаков с конкретными признаками. Таким образом извлечение лексических и метроритмических признаков, характерных для жанра и стиля и их комбинаций улучшает процесс автоматизированной обработки текстов на русском языке, что продемонстрировано на основе подготовленного корпуса поэтических текстов А.С. Пушкина и К.Н. Батюшкова. Полученные результаты могут быть использованы для облегчения верификации классификатора, а также составления списка характерных для жанра и стиля признаков в творчестве того или иного поэта.

Об авторах:

Барахнин Владимир Борисович, доктор технических наук, доцент, ведущий научный сотрудник, Институт вычислительных технологий Сибирского отделения Российской академии наук (630090, Россия, г. Новосибирск, пр. Академика Лаврентьева, д. 6); профессор, кафедра общей информатики, факультет информационных технологий, Новосибирский национальный исследовательский государственный университет (630090, Россия, г. Новосибирск, ул. Пирогова, д. 1), ORCID: http://orcid.org/0000-0003-3299-0507, bar@ict.nsc.ru

Кожемякина Ольга Юрьевна, кандидат филологических наук, старший научный сотрудник, Институт вычислительных технологий Сибирского отделения Российской академии наук (630090, Россия, г. Новосибирск, пр. Академика Лаврентьева, д. 6), ORCID: http://orcid.org/0000-0003-3619-1120, olgakozhemyakina@mail.ru

Рычкова Елена Владимировна, кандидат физико-математических наук, доцент, научный сотрудник, Институт вычислительных технологий Сибирского отделения Российской академии наук (630090, Россия, г. Новосибирск, пр. Академика Лаврентьева, д. 6); доцент кафедры массовых коммуникаций, Гуманитарный институт, Новосибирский национальный исследовательский государственный университет (630090, Россия, г. Новосибирск, ул. Пирогова, д. 1), ORCID: http://orcid.org/0000-0001-9692-8441, helen@ict.nsc.ru

Пастушков Илья Сергеевич, аспирант, Институт вычислительных технологий Сибирского отделения Российской академии наук (630090, Россия, г. Новосибирск, пр. Академика Лаврентьева, д. 6), ORCID: http://orcid.org/0000-0002-0341-7931, pas2shkov.ilya@gmail.com

Борзилова Юлия Сергеевна, аспирант, Институт вычислительных технологий Сибирского отделения Российской академии наук (630090, Россия, г. Новосибирск, пр. Академика Лаврентьева, д. 6), ORCID: http://orcid.org/0000-0002-8265-9356, borzilova.ys@gmail.com

Keywords

Patterns recognition; principal component analysis; automated analysis of poetic texts; algorithm of classification; ensembling.

Abstract

This paper describes the algorithm of automatic extraction of the characteristic features for the genre and the style. This work was carried out in the framework of the development of a software system created in the Institute of Computational Technologies of SB RAS and designed for a complex analysis of metrorhythmic and genre-stylistic characteristics of poetic texts in Russian. The paper presents the structure of the software system developed in the ICT SB RAS and intended for a complex analysis of metrorhythmic and genre-stylistic characteristics of poetic texts in Russian. The system organically combines both original program modules which are created directly by the system developers and intended for the solution of the single-purpose tasks of the analysis of the poetic texts, and open access software products. The generalized approach, which allows to consider the poetic features in the form of a vector, on the one hand, allows to use the modern algorithms of the classification and their ensembles, on the other, such approach has the disadvantages for small volumes of the texts with which it is necessary to work. Therefore, the presence of such a step as verification allows the specialists to adjust the operation of the system based on an expert knowledge, and also makes the classification process transparent. As a tool, the Python libraries were used: scikit-learn, in which the algorithms of the classification and also the methods of their combination were implemented; and ELI5, which allows to establish a correspondence between the components of the feature vector with specific features. So, the extraction of lexical and metrorhythmic features which are characteristic for the genre and style and of their combinations improved the process of automated processing of poetic texts in Russian what is shown on the base of the corpus of poetic texts of A.S. Pushkin and K.N. Batyushkov. The obtained results can be used for the verification of the classifier and for a list of characteristic features for the genre and the style of a poet.

Введение

Исследования по извлечению характерных для жанра и стиля лексических и метроритмических признаков и их сочетаний проводятся в настоящее время на базе Института вычислительных технологий СО РАН (г. Новосибирск). Работа выполняется в рамках общего проекта по автоматизации работы по проведению анализа поэтических текстов (рис. 1).

Рис. 1. Общая схема проекта ИВТ СО РАН по автоматизации работы для проведения анализа поэтических текстов

Fig. 1. General scheme of the ICT SB RAS project on automation of analyzing poetic texts

Представленная схема наглядно показывает, как конкретные разработки каждого уровня анализа, так их взаимосвязь и взаимозависимость в рамках общей постановки задачи. Целью проекта является пакетная обработка поэтических текстов, включающая в себя несколько равноценных компонентов, в виде подзадач. Разработка компонентов системы не последовательна, мы не выделяем приоритетных задач, поскольку все компоненты проекта взаимосвязаны в рамках получения результата. Проект состоит из следующих компонентов:

- инициализация системы для пакетной обработки текстов;
- структурный анализ;
- семантический анализ;
- прагматический анализ;
- разработка систем взаимодействия с конечным пользователем.

После инициализации системы и загрузки корпусов текстов, в задачах следующего компонента осуществляется структурный анализ метроритмических характеристик, таких как:

- метрика стихотворения и стопность;
- количество строк, без учета пустых;
- рифмовка строфики;
- количество мужских окончаний последних слов в стихотворных строках;
- количество женских окончаний последних слов в стихотворных строках;
- количество дактилических и др. окончаний последних слов в стихотворных строках;
- количество нерифмованных мужских окончаний;
- количество нерифмованных женских окончаний;
- количество нерифмованных дактилических и других окончаний;
- количество строк без конечных слов;
- тип строфической формы.

Кроме того, определяется фактура стиха (термин «фактура», введенный Г. Шенгели [1] и используемый участниками

проектной группы в исследованиях, получает терминологическое обоснование). Анализ метроритмики стиха реализован в системе [2] в виде веб-приложения.

Компонент семантического анализа включает в себя извлечение словосочетаний и характеристик слов, связанных с синонимией: эти процессы тесно связаны между собой и взаимозависимы. Компонент разработок связан с прагматическими аспектами анализа поэтических текстов [3, 4]. Определение жанра, стиля и других характеристик прагматического уровня текста выполняется с использованием машинного обучения [5], что позволяет построить рекомендательную систему на основе интеллектуального анализа лингвистики (ИАЛ) [6]. На этапе взаимодействия с конечным пользователем для этой системы создается интерфейс для работы филологов (рис. 2).

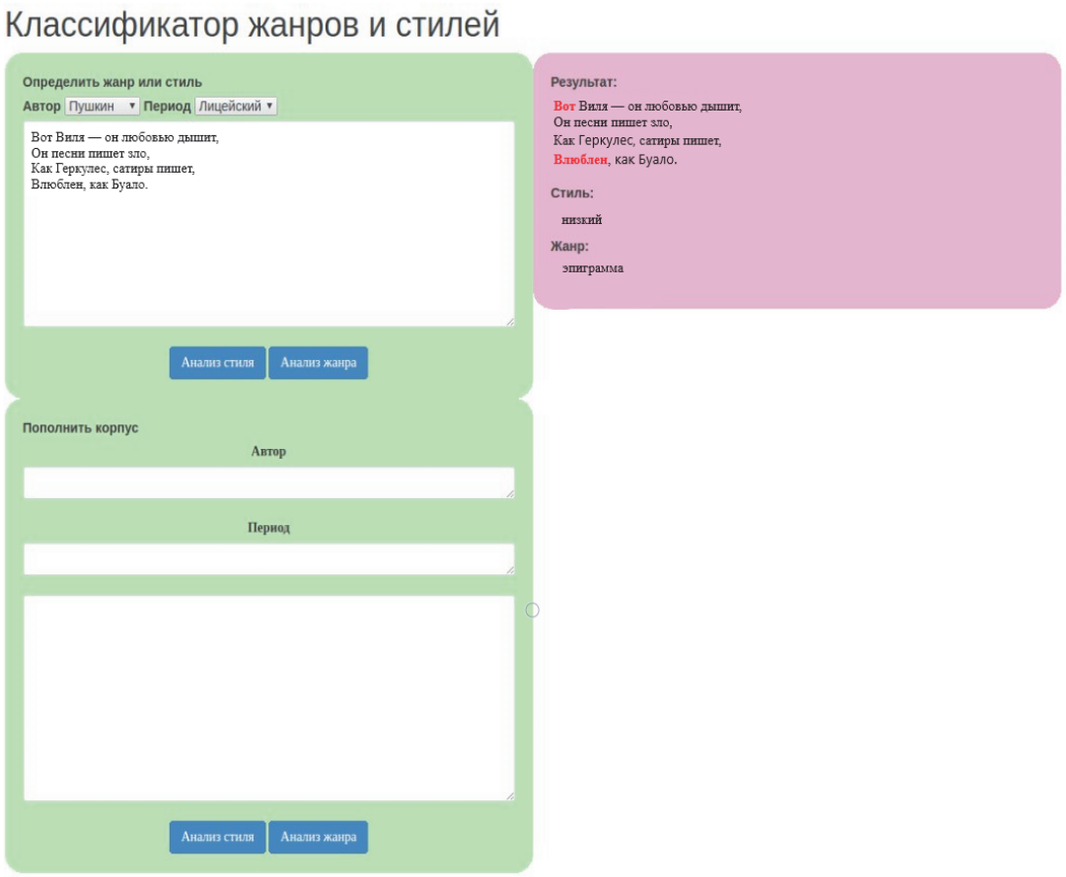

Рис. 2. Интерфейс системы для определения жанра и стиля с подсказкой ключевых слов
Fig. 2. System Interface for Genre and Style Specification with Keyword Tips

Этот интерфейс включает в себя пакетную загрузку текстов в хранилище данных и получение выходных параметров поэтических текстов после обработки. В качестве входных данных используются корпуса текстов, предварительно размеченные экспертом и размещенные в системе. С помощью блока «Пополнить корпус» пользователь имеет возможность внести текст с указанием автора и периода анализируемого материала. Отдельно определяются стиль и жанр. В отдельном блоке отображается текст с выделением ключевых слов, на основании которых система сделала вывод о принадлежности к тому или иному стилю и жанру.

В общем случае работа с поэтическими текстами проходит следующие этапы:

Этап 1. Пакетная выгрузка поэтических текстов из базы данных. Анализ с помощью системы [2]. Полученная метроритмическая статистика выгружается обратно в базу данных.

Этап 2. Автоматическое извлечение словосочетаний с учетом поэтической синонимии. Выгрузка полученных результатов в базу данных.

Этап 3. Определение жанрово-стилевых характеристик с применением методов машинного обучения. Выгрузка результатов в базу.

Этап 4. Организация работы с полученными характеристиками с возможностью их сравнения.

В настоящей работе мы предлагаем и используем дополнительные характеристики для разработки совместного («двумерного») классификатора жанровых типов и стилистической окраски поэтических текстов [7, 8]. Наша цель — это не создание новых теорий жанрово-стилевых отношений в рамках литературных произведений, а разработка системы анализа, позволяющей корректно соотносить стилевую окраску текста с его жанровой принадлежностью, что актуально для исследователей в области информатики, потому что мы говорим об инструментах, используемых не только в строго лингвистических исследованиях [9 – 13].

Обученный классификатор позволяет выделить ключевые для определения жанра и стиля лексические признаки исходя из весовых коэффициентов классификатора в совокупности с его качеством. Для анализа признаков применяется стандартный подход — анализ главных компонент (principal component analysis, PCA), который выделяет наиболее значимые компоненты в матрице, исходя из числа сингулярности. Так как размерность вектора признаков приводит к большим затратам машинного времени, то в процессе PCA выполняется уменьшение размерности признаков, используя сингулярное разложение (singular value decomposition, SVD).

Постановка задачи и применяемый инструментарий

Анализ уже существующей информации был и остается актуальной проблемой для исследователей [14 – 18]. В задачах обработки поэтических текстов [19, 20], к анализу информации можно отнести классификацию по жанру и стилю. Однако, даже при достаточном качестве классификатора, сложно осуществить его дальнейшее развитие в ситуации, когда матрица признаков является обезличенной, иначе говоря, при отсутствии прямой связи между компонентами векторов признаков и самими признаками. Извлечение наиболее важных признаков для классификации позволяет помечать элементы текста, весовые коэффициенты которых можно будет изменить. Это позволит: во-первых, сделать работу классификатора прозрачной для специалистов без углубленной математической подготовки; во-вторых, преодолеть порог качества работы классификатора, который был обусловлен генерализованным подходом к классификации. Активное внедрение информационных технологий в лингвистику значительно упрощает процесс верификации характерных для жанра и стиля признаков [21 – 23].

Извлечение наиболее важных признаков для классификации состоит из анализа весовых коэффициентов обученного классификатора и оценок точности и полноты для конкретных случаев.

В качестве инструмента были применены библиотеки языка Python: scikit-learn [24], в которой реализованы алгоритмы классификации, а также методы их комбинирования, и ELI5 [25], позволяющая установить соответствие между компонентами вектора признаков с конкретными признаками. В табл. 1 приведены выходные данные, полученные из ELI5.

Таблица 1. Характеристики весов для каждого стиля
Table 1. Weight characteristics for each style

Низкий стиль Признаки с наибольшим весом		**Средний стиль** Признаки с наибольшим весом		**Высокий стиль** Признаки с наибольшим весом	
Вес	**Признак**	**Вес**	**Признак**	**Вес**	**Признак**
+ 1.991	ты	+ 1.702	он	+ 1.522	из словаря русского языка XVIII в.
+ 1.925	которого	+ 0.702	друг	+ 1.518	роковой
10174 признака с положительным весом		*11710 признаков с положительным весом*		*15007 признаков с положительным весом*	
25605 признаков с отрицательным весом		*24069 признаков с отрицательным весом*		*20772 признака с отрицательным весом*	
– 1.686	вас	– 1.379	<BIAS>	– 1.764	которого
– 10.453	<BIAS>			– 2.171	чего
				– 5.013	<BIAS>

Столбцы таблицы соответствуют целевым классам. В каждом столбце приведены характеристики и их веса. Байесовский признак пересечения показан как <BIAS>. Положительные / отрицательные веса выделены двумя цветами. Мы можем рассмотреть признаки и их веса, т.к. мы используем модель «мешок слов», либо вектор n-грамм и классификатор (таким образом имеется прямое соответствие между отдельными словами или n-граммами и коэффициентами классификатора). Сложность вызывают другие признаки, используемые в классификаторе. Например, некоторые нелексические признаки, такие как наличие слов из словаря русского языка XVIII в. и особых метроритмических комбинаций.

Алгоритм извлечения признаков и полученные результаты

Нами реализован совместный («двумерный») классификатор жанровых типов и стилистической окраски поэтических текстов, считая, таким образом, что классификатор сам по себе — многомерная структура, основанная на множестве параметров, которые определяют предметную область. Многомерные классификаторы, связанные с такими сложными (для однозначного определения) категориями как жанр и стиль требуют поэтапной разработки каждого параметра анализа, чтобы исключить возможные ошибки и вариативность результатов. Для анализа использовалась лирика А.С. Пушкина 1818 – 1825 годов и вся лирика К.Н. Батюшкова.

Все корпуса текстов были разделены на три части: лирика А.С. Пушкина лицейского периода, лирика А.С. Пушкина 1818 – 1825 годов и вся лирика К.Н. Батюшкова. Для каждого стихотворения мы извлекаем несколько признаков, включая признаки TF-IDF матрицы: вхождение слов из стихотворения в словарь русского языка XVIII в., метроритмические признаки (тип рифмы, размер, строфа, количество мужских и женских окончаний, год написания стиха). Использование древнерусской и церковно-славянской лексики было определено разницей между упомянутым выше словарем и корпусом русской Википедии и декодировалось как вектор с количеством использования для каждого из этих слов.

Для решения проблемы разности весов множеств, принадлежащих к тем или иным классам, был использован алгоритм SMOTE («метод роя частиц») а также случайное дублирование примеров класса меньшей мощности. После извлечения признаков логистическая регрессия была обучена и признаки, оказавшиеся наиболее важными для принятия решения, были использованы для обучения модели многослойного персептрона [26]. Полученная средняя f1-мера в ~95% по результатам перекрёстной проверки на трёх разбиениях для каждого автора, однако следует заметить, что некоторые классы определяются хуже, так как представлены слишком малым количеством примеров.

Результаты экспериментов: среднее значение, максимум, минимум f1-меры правильных предсказаний метода на перекрёстной проверке представлены в табл. 2. Алгоритм был реализован на языке программирования Python с использованием библиотеки scikit-learn.

Таблица 2. Эксперимент с определением жанра/стиля на многослойном персептроне с признаками, которые оказались наилучшими для логистической регрессии
Table 2. Genre / style determination experiment on a multi-layer perceptron with features that were best for logistic regression

	Средняя f1-мера	**Максимум f1-меры**	**Минимум f1-меры**
А.С. Пушкин	0.93	0.95	0.92
К.Н. Батюшков	0.91	0.92	0.89

Для уменьшения размерности вектора признаков были использованы усечённый вариант SVD (singular value decomposition, разложение сингулярных значений) и SVD как часть PCA (principal component analysis, анализ главных компонент). PCA используется для разложения многообразного набора данных в набор последовательных ортогональных компонент, объясняющих максимальный объём вариативности. Это позволяет проецировать данные в сингулярное пространство с масштабированием каждой компоненты по области значений. Это часто полезно в моделях последовательной обработки для уверенных предположений о подобности объектов, например, в случае машины опорных векторов с ядром в виде радиальной базовой функции, а также метода кластеризации *k*-средних.

Усечённый SVD реализует вариант разложения сингулярных значений, считая только *k* наибольших сингулярных значений по заданному параметру *k*.

При применении SVD к матрицам терминов-документов, например, матрица TF-IDF, получается преобразование более известное как латентный семантический анализ (LSA) так как он преобразует матрицы такого рода в семантическое пространство низкой размерности. В частности, LSA устойчив к проблемам синонимии и полисемии (оба термина подразумевают многозначность слов), которые в случае матриц терминов-документов приводят к их чрезмерной разреженности и плохим результатам таких метрик близости как косинусное расстояние.

Заключение

В работе изложена структура программной системы, разработанной в ИВТ СО РАН и предназначенной для комплексного анализа метроритмических и жанрово-стилистических характеристик поэтических текстов на русском языке. Система органично сочетает в себе как оригинальные программные модули, созданные непосредственно разработчиками системы и предназначенные для решения узкоспециализированных задач анализа поэтических текстов, так и программные продукты открытого доступа. Обобщённый подход, позволяющий рассматривать поэтические признаки в виде вектора, с одной стороны, позволяет использовать современные алгоритмы классификации и их ансамбли, с другой, такой подход имеет недостатки при малых объёмах корпусов, с которыми приходится работать. Поэтому наличие такого шага как верификация позволяет специалистам корректировать работу системы исходя из экспертных знаний, а также делает процесс классификации прозрачным. Достигнутое среднее значение f1-меры примерно равно 95 % при перекрестной проверке на трех разбиениях для каждого автора, однако при дополнительной верификации это значение уменьшается на 2 %. Таким образом извлечение лексических и метроритмических признаков, характерных для жанра и стиля и их комбинаций улучшает процесс автоматизированной обработки поэтических текстов на русском языке.

Благодарность

Работа выполнена в рамках темы государственного задания № АААА-А17-117120670141-7 (№ 0316-2018-0009) при частичной поддержке гранта Российского фонда фундаментальных исследований № 18-07-01457 и гранта Министерства образования и науки Республики Казахстан № BR05236839.

Acknowledgement

The work was carried out as part of the state assignment number AAAA-A17-117120670141-7 (No. 0316-2018-0009) with partial support from the Russian Foundation for Basic Research grant No. 18-07-01457 and the grant from the Ministry of Education and Science of the Republic of Kazakhstan No. BR05236839.

Список использованных источников

[1] *Шенгели Г.А.* Техника стиха. М.: ГИХЛ, 1960. 312 с. URL: http://imwerden.de/pdf/shengeli_tekhnika_stikha_1960_text.pdf (дата обращения: 25.09.2018).

[2] Анализ поэтических текстов онлайн. [Электронный ресурс]. URL: http://poem.ict.nsc.ru/ (дата обращения: 25.09.2018).

[3] *Barakhnin V., Kozhemyakina O.* About the automation of the complex analysis of Russian poetic text // CEUR Workshop Proceedings. 2012. Vol. 934. Pp. 167-171. URL: http://ceur-ws.org/Vol-934/paper27.pdf (дата обращения: 25.09.2018).

[4] *Мельчук И.А.* Язык: от смысла к тексту. М.: Языки славян-

ских культур, 2012. 190 с. URL: http://biblioclub.ru/index.php?page=book&id=219899 (дата обращения: 25.09.2018).

[5] *Friedman J.H.* Stochastic Gradient Boosting // Computational Statistics and Data Analysis. 2002. Vol. 38, issue 4. Pp. 367-378. DOI: 10.1016/S0167-9473(01)00065-2

[6] *Korobov M.* Morphological Analyzer and Generator for Russian and Ukrainian Languages // Communications in Computer and Information Science. 2015. Vol. 542. Pp. 320-332. URL: https://elibrary.ru/item.asp?id=26927893 (дата обращения: 25.09.2018).

[7] *Barakhnin V.B., Kozhemyakina O.Yu., Pastushkov I.S.* Comparative analysis of methods of automated classification of poetic texts based on lexical signs // CEUR Workshop Proceedings. 2017. Vol. 2022. Pp. 252-257. URL: http://ceur-ws.org/Vol-2022/paper41.pdf (дата обращения: 25.09.2018).

[8] *Barakhnin V.B., Kozhemyakina O.Yu., Pastushkov I.S.* Automated Determination of the Type of Genre and Stylistic Coloring of Russian Texts // ITM Web of Conferences. 2017. Vol. 10. Art. 02001. DOI: https://doi.org/10.1051/itmconf/20171002001

[9] *Barakhnin V.B., Fedotov A.M., Bakiyeva A.V., Bakiyev M.N., Tazhibayeva S.Zh., Batura T.V., Kozhemyakina O.Yu., Tussupov D.A., Sambetbaiyeva M.A., Lukpanova L.Kh.* The Software System for the Study the Morphology of the Kazakh Language // The European Proceedings of Social & Behavioural Sciences. 2017. Vol. XXXIII. Pp.18-27. URL: http://www.futureacademy.org.uk/files/images/upload/ICPE2017F3.pdf (дата обращения: 25.09.2018).

[10] *Barakhnin V.B., Kozhemyakina O.Yu., Zabaykin A.V.* Usage of modern computer technologies in the learning process of the philologists of complex analysis of Russian poetic texts // SHS Web of Conferences. 2016. Vol. 29. Art.UNSP02002. DOI: http://dx.doi.org/10.1051/shsconf/20162902002

[11] *Барахнин В.Б., Кожемякина О.Ю., Забайкин А.В.* Алгоритмы комплексного анализа русских поэтических текстов с целью автоматизации процесса создания метрических справочников и конкордансов // CEUR Workshop Proceedings. 2015. Vol. 1536. Pp. 138-143. URL: http://ceur-ws.org/Vol-1536/paper21.pdf (дата обращения: 25.09.2018).

[12] *Mansurova M.E., Barakhnin V.B., Aubakirov S.S., Khibatkhanuly Ye., Mussina A.B.* Parallel text document clustering based on genetic algorithm // CEUR Workshop Proceedings. 2017. Vol. 1839. Pp. 218-232. URL: http://ceur-ws.org/Vol-1839/MIT2016-p20.pdf (дата обращения: 25.09.2018).

[13] *Мансурова М.Е., Барахнин В.Б., Аубакиров С.С., Хибатханулы Е., Мусина А.Б.* Разработка параллельного алгоритма кластеризации текстовых документов FRiS-Tax на основе технологии MPI // CEUR Workshop Proceedings. 2016. Vol. 1576. Pp. 244-256. URL: http://ceur-ws.org/Vol-1576/128.pdf (дата обращения: 25.09.2018).

[14] *Chawla N.V.* Data Mining for Imbalanced Datasets: An Overview // Data Mining and Knowledge Discovery Handbook. Springer-Verlag, 2010. Pp. 875-886. DOI: 10.1007/978-0-387-09823-4_45

[15] *Mikolov T., Kai Chen, Corrado G., Dean J.* Efficient Estimation of Word Representations in Vector Space // Computation and Language. 2013. URL: https://arxiv.org/pdf/1301.3781.pdf (дата обращения: 25.09.2018).

[16] *Москвина А.Д., Митрофанова О.А., Ерофеева А.Р., Харабет Я.К.* Автоматическое выделение ключевых слов и словосочетаний из русскоязычных корпусов текстов с помощью алгоритма RAKE // Труды международной конференции «Корпусная лингвистика-2017». СПб., 2017. С. 268-274. URL: https://elibrary.ru/item.asp?id=32425675 (дата обращения: 25.09.2018).

[17] *Соколова Е.В., Митрофанова О.А.* Автоматическое извлечение ключевых слов и словосочетаний из русскоязычных текстов с помощью алгоритма KEA // Компьютерная лингвистика и вычислительные онтологии. Вып. 1. Труды XX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2017). СПб.: Университет ИТМО, 2017. С. 157-165. URL: http://openbooks.ifmo.ru/ru/file/6522/6522.pdf (дата обращения: 25.09.2018).

[18] *Москвина А.Д., Орлова Д., Паничева П.В., Митрофанова О.А.* Разработка ядра синтаксического анализатора для русского языка на основе библиотек NLTK // Компьютерная лингвистика и вычислительные онтологии. Труды XIX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2016). СПб.: Университет ИТМО, 2016. С. 44-45. URL: http://openbooks.ifmo.ru/ru/file/4103/4103.pdf (дата обращения: 25.09.2018).

[19] *Vetulani Z., Obrębski T., Vetulani G.* Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora // Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference. 2007. Pp. 267-268. URL: http://www.aaai.org/Papers/FLAIRS/2007/Flairs07-055.pdf (дата обращения: 25.09.2018).

[20] *Кощеева С.С.* Сравнение методов автоматического выделения глагольно-именных словосочетаний // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2014). СПб.: Университет ИТМО, 2014. С. 298-303. URL: http://ojs.ifmo.ru/index.php/IMS/article/view/270/266 (дата обращения: 25.09.2018).

[21] *Ванюшкин А.С., Гращенко Л.А.* Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. 2016. № 19. С. 85-93. URL: https://elibrary.ru/item.asp?id=25864034 (дата обращения: 25.09.2018).

[22] *Орлов Ю.Н., Осминин К.П.* Определение жанра и автора литературного произведения статистическими методами // Прикладная информатика. 2010. № 2(26). С. 95-108. URL: https://elibrary.ru/item.asp?id=13569170 (дата обращения: 25.09.2018).

[23] *Орлов Ю.Н., Осминин К.П.* Методы статистического анализа литературных текстов. М.: URSS, 2017. 312 с.

[24] Scikit-learn: Machine Learning in Python. [Электронный ресурс]. URL: http://scikit-learn.org/ (дата обращения: 25.09.2018).

[25] ELI5's documentation. [Электронный ресурс]. URL: http://eli5.readthedocs.io/ (дата обращения: 25.09.2018).

[26] *Haykin S.* Neural Networks: A Comprehensive Foundation. 2nd Edition. Prentice Hall PTR Upper Saddle River, NJ, USA, 1998. 842 p.

Поступила 25.09.2018; принята в печать 20.10.2018;
опубликована онлайн 10.12.2018.

References

[1] Shengeli G.A. The technique of a verse. Moscow: GIHL, 1960. 312 p. Available at: http://imwerden.de/pdf/shengeli_tekhnika_stikha_1960_text.pdf (accessed 25.09.2018). (In Russian)

[2] Analysis of the poetic texts online. Available at: http://poem.ict.nsc.ru/ (accessed 25.09.2018). (In Russian)

[3] Barakhnin V., Kozhemyakina O. About the automation of the complex analysis of Russian poetic text. *CEUR Workshop Proceedings*. 2012; 934:167-171. Available at: http://ceur-ws.org/Vol-934/paper27.pdf (accessed 25.09.2018).

[4] Melchuk I.A. Language: from meaning to text. M.: Yazyki slavyanskikh kultur, 2012. 190 p. Available at: http://biblioclub.ru/index.php?page=book&id=219899 (accessed 25.09.2018). (In Russian)

[5] Friedman J.H. Stochastic Gradient Boosting. *Computational Statistics and Data Analysis*. 2002; 38(4):367-378. DOI: 10.1016/S0167-9473(01)00065-2

[6] Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. *Communications in Computer and Information Science*. 2015; 542:320-332. Available at: https://elibrary.ru/item.asp?id=26927893 (accessed 25.09.2018).

[7] Barakhnin V.B., Kozhemyakina O.Yu., Pastushkov I.S. Comparative analysis of methods of automated classification of poetic texts based on lexical signs. *CEUR Workshop Proceedings.* 2017; 2022:252-257. Available at: http://ceur-ws.org/Vol-2022/paper41.pdf (accessed 25.09.2018).

[8] Barakhnin V.B., Kozhemyakina O.Yu., Pastushkov I.S. Automated Determination of the Type of Genre and Stylistic Coloring of Russian Texts. *ITM Web of Conferences*. 2017; 10:02001. DOI: https://doi.org/10.1051/itmconf/20171002001

[9] Barakhnin V.B., Fedotov A.M., Bakiyeva A.V., Bakiyev M.N., Tazhibayeva S.Zh., Batura T.V., Kozhemyakina O.Yu., Tussupov D.A., Sambetbaiyeva M.A., Lukpanova L.Kh. The Software System for the Study the Morphology of the Kazakh Language. *The European Proceedings of Social & Behavioural Sciences*. 2017; XXXIII:18-27. Available at: http://www.futureacademy.org.uk/files/images/upload/ICPE2017F3.pdf (accessed 25.09.2018).

[10] Barakhnin V.B., Kozhemyakina O.Yu., Zabaykin A.V. Usage of modern computer technologies in the learning process of the philologists of complex analysis of Russian poetic texts. *SHS Web of Conferences*. 2016; 29:UNSP02002. DOI: http://dx.doi.org/10.1051/shsconf/20162902002

[11] Barakhnin V.B., Kozhemyakina O.Y., Zabaykin A.V. The Algorithms of Complex Analysis of Russian Poetic Texts for the Purpose of Automation of the Process of Creation of Metric Reference Books and Concordances. *CEUR Workshop Proceedings.* 2015; 1536:138-143. Available at: http://ceur-ws.org/Vol-1536/paper21.pdf (accessed 25.09.2018). (In Russian)

[12] Mansurova M.E., Barakhnin V.B., Aubakirov S.S., Khibatkhanuly Ye., Mussina A.B. Parallel text document clustering based on genetic algorithm. *CEUR Workshop Proceedings*. 2017; 1839:218-232. Available at: http://ceur-ws.org/Vol-1839/MIT2016-p20.pdf (accessed 25.09.2018).

[13] Mansurova M.E., Barakhnin V.B., Aubakirov S.S., Khibatkhanuly Ye., Mussina A.B. Development of parallel FRiS-Tax text document clustering algorithm based on MPI technology. *CEUR Workshop Proceedings*. 2016; 1576:244-256. Available at: http://ceur-ws.org/Vol-1576/128.pdf (accessed 25.09.2018). (In Russian)

[14] Chawla N.V. Data Mining for Imbalanced Datasets: An Overview. *Data Mining and Knowledge Discovery Handbook*. Springer-Verlag, 2010, pp. 875-886. DOI: 10.1007/978-0-387-09823-4_45

[15] Mikolov T., Kai Chen, Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. *Computation and Language*. 2013. Available at: https://arxiv.org/pdf/1301.3781.pdf (accessed 25.09.2018).

[16] Moskvina A.D., Mitrofanova O.A., Erofeeva A.R., Charabet Ja.K. Automatic extraction of key words and phrases from Russian text corpora by means of RAKE algorithm. *Proceedings of the International conference "Corpora linguistics-2017"*. 2017, pp. 268-274. Available at: https://elibrary.ru/item.asp?id=32425675 (accessed 25.09.2018). (In Russian)

[17] Sokolova E., Mitrofanova O. Automatic Keyphrase Extraction by applying KEA to Russian texts. *Proceedings of the XX International joint conference "Internet and Modern Society" (IMS-2017).* 2017, pp. 157-165. Available at: http://openbooks.ifmo.ru/ru/file/6522/6522.pdf (accessed 25.09.2018). (In Russian)

[18] Moskvina A., Orlova D., Panicheva P., Mitrofanova O. Development of the Core for Syntactic Parser for Russian based on NLTK libraries. *Proceedings of the XIX International joint conference "Internet and Modern Society" (IMS-2016)*. 2016, pp. 44-45. Available at: http://openbooks.ifmo.ru/ru/file/4103/4103.pdf (accessed 25.09.2018). (In Russian)

[19] Vetulani Z., Obrębski T., Vetulani G. Towards a Lexicon-Grammar of Polish: Extraction of Verbo-Nominal Collocations from Corpora. *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society Conference*. 2007, pp. 267-268. Available at: http://www.aaai.org/Papers/FLAIRS/2007/Flairs07-055.pdf (accessed 25.09.2018).

[20] Koshcheeva S. Comparing Methods of Automatic Verb-Noun Collocation Extraction. *Proceedings of the XVII all-Russian joint conference "Internet and Modern Society" (IMS-2014)*. 2014, pp. 298-303. Available at: http://ojs.ifmo.ru/index.php/IMS/article/view/270/266 (accessed 25.09.2018). (In Russian)

[21] Vanyushkin A.S., Grashchenko L.A. Methods and algorithms of the extraction of the keywords. *New information technologies in automated systems.* 2016; 19:85-93. Available at: https://elibrary.ru/item.asp?id=25864034 (accessed 25.09.2018). (In Russian)

[22] Orlov Yu.N., Osminin K.P. Determination of the genre and author of the literary work by statistical methods. *Journal of Applied Informatics.* 2010; 2(26):95-108. Available at: https://elibrary.ru/item.asp?id=13569170 (accessed 25.09.2018). (In Russian)

[23] Orlov Yu.N., Osminin K.P. Methods of statistical analysis of literary texts. M.: URSS, 2012. 312 p. (In Russian)

[24] Scikit-learn: Machine Learning in Python. Available at: http://scikit-learn.org/ (accessed 25.09.2018).

[25] ELI5's documentation. Available at: http://eli5.readthedocs.io/ (accessed 25.09.2018).

[26] Haykin S. Neural Networks: A Comprehensive Foundation. 2nd ed. Prentice Hall PTR Upper Saddle River, NJ, USA, 1998. 842 p.

Submitted 25.09.2018; revised 20.10.2018;
published online 10.12.2018.

About the authors:

Vladimir B. Barakhnin, Doctor of Technical Sciences, Associate Professor, Lead researcher, Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences (6 Ac. Lavrentieva Av., Novosibirsk 630090, Russia); Professor, Department of Information Technologies, Novosibirsk State University (1, Pirogova Str., Novosibirsk 630090, Russia), ORCID: http://orcid.org/0000-0003-3299-0507, bar@ict.nsc.ru
Olga Yu. Kozhemyakina, Candidate of Philology, Senior Researcher, Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences (6 Ac. Lavrentieva Av., Novosibirsk 630090, Russia), ORCID: http://orcid.org/0000-0003-3619-1120, olgakozhemyakina@mail.ru
Elena V. Rychkova, Candidate of Physical and Mathematical Sciences, Associate Professor, Researcher, Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences (6 Ac. Lavrentieva Av., Novosibirsk 630090, Russia); Associate Professor, Institute for the Humanities, Novosibirsk State University (1 Pirogova Str., Novosibirsk 630090, Russia), ORCID: http://orcid.org/0000-0001-9692-8441, helen@ict.nsc.ru
Ilya S. Pastushkov, post-graduate student, Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences (6 Ac. Lavrentieva Av., Novosibirsk 630090, Russia), ORCID: http://orcid.org/0000-0002-0341-7931, pas2shkov.ilya@gmail.com
Yuliya S. Borzilova, post-graduate student, Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Sciences (6 Ac. Lavrentieva Av., Novosibirsk 630090, Russia), ORCID: http://orcid.org/0000-0002-8265-9356, borzilova.ys@gmail.com